

Left-corner Methods for Syntactic Modeling with Universal Structural Constraints

A dissertation
submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in
Informatics
of SOKENDAI
Graduate University for Advanced Studies

Hiroshi Noji
June 2016

© Copyright by Hiroshi Noji 2016
All Rights Reserved

Abstract

Explaining the syntactic variation and universals including the constraints on that variation across languages in the world is essential both from a theoretical and practical point of view. It is in fact one of the main goals in linguistics. In computational linguistics, these kinds of syntactic regularities and constraints could be utilized as prior knowledge about grammars, which would be valuable for improving the performance of various syntax-oriented systems such as parsers or grammar induction systems. This thesis is about such syntactic universals.

The primary goal in this thesis is to identify better syntactic constraint or bias, that is language independent but also efficiently exploitable during sentence processing. We focus on a particular syntactic construction called center-embedding, which is well studied in psycholinguistics and noted to cause particular difficulty for comprehension. Since people use language as a tool for communication, one expects such complex constructions to be avoided for communication efficiency. From a computational perspective, center-embedding is closely relevant to a *left-corner* parsing algorithm, which can capture the degree of center-embedding of a parse tree being constructed. This connection suggests left-corner methods can be a tool to exploit the universal syntactic constraint that people avoid generating center-embedded structures. We explore such utilities of center-embedding as well as left-corner methods extensively through several theoretical and empirical examinations.

We base our analysis on dependency syntax. This is because our focus in this thesis is the language universality. Now the number of available dependency treebanks are growing rapidly compared to the treebanks of phrase-structure grammars thanks to the recent standardization efforts of dependency treebanks across languages, such as the Universal Dependencies project. We use these resources, consisting of more than 20 treebanks, which enable us to examine the universality of particular language phenomena, as we pursue in this thesis.

First, we quantitatively examine the universality of center-embedding avoidance using a collection of dependency treebanks. Previous studies on center-embedding in psycholinguistics have been limited to behavioral studies focusing on particular languages or sentences. Our study contrasts with these previous studies, and provides the first quantitative results on center-embedding avoidance. Along with these experiments, we provide a parser that can capture the degree of center-embedding of a *dependency* tree being built, by extending a left-corner parsing algorithm for dependency grammars. The main empirical finding in this study is that center-embedding is in fact a rare phenomenon across languages. This result also suggests a left-corner parser could be utilized as a tool exploiting the universal syntactic constraints in languages.

We then explore such utility of a left-corner parser in the application of unsupervised grammar induction. In this task, the input to the algorithm is a collection of sentences, from which the

model tries to extract the salient patterns on them as a grammar. This is a particularly hard problem although we expect the universal constraint may help in improving the performance since it can effectively restrict the possible search space for the model. We build the model by extending the left-corner parsing algorithm for efficiently tabulating the search space except those involving center-embedding up to a specific degree. Again, we examine the effectiveness of our approach on many treebanks, and demonstrate that often our constraint leads to better parsing performance. We thus conclude that left-corner methods are particularly useful for syntax-oriented systems, as it can exploit efficiently the inherent universal constraints in languages.

Acknowledgments

I could not have finished writing this thesis without the support of many people.

I am greatly indebted to my advisor, Professor Yusuke Miyao for his continuous help and support throughout my PhD. He was always giving me many constructive and encouraging suggestions when I had trouble in writing a paper, preparing presentation slides, interpreting experimental results, and so on, which cannot be listed in this space. He is arguably the best advisor I have met so far, and also is an excellent research role model for me.

I would like to thank my thesis committee, Professor Hiroshi Nakagawa, Professor Edson Miyamoto, Professor Makoto Kanazawa, and Professor Daichi Mochihashi. They all have different backgrounds, and gave me very valuable and insightful comments from their own perspectives. Professor Daichi Mochihashi guided me in my masters into the research field of unsupervised learning, which was my starting point of this work focusing on unsupervised grammar induction.

My first advisor in my masters in University of Tokyo, Professor Kumiko Tanaka-Ishii, introduced me to the research field of computational linguistics. Though I could not interact with her in the last three years in my PhD, her advices to me had great impacts on my thinking on research. The fundamental idea behind this thesis, exploring the universal property of language from a computational perspective, was arguably the one inspired by her philosophy on computational linguistics.

I am grateful to Mark Johnson for hosting me in total three times (!) as a visitor in Macquarie University. Regular meetings with him were always exciting, and have sharpen my thinking on the tasks of unsupervised learning. It was really my fortunate to have collaboration with a great researcher as you during my PhD.

As a member in Miyao lab, I had opportunities to interact with many great researchers in NII. Special thanks to Takuya Matsuzaki, who was always welcoming informal discussion in his room, and I remember that the core research idea on this thesis, left-corner parsing, has been came up with during a conversation with him. Pascual Martínez-Gómez, Yuka Tateishi, and Sumire Uematsu always gave me many constructive comments especially in my practices on presentations. I am grateful to Bevan Johns, who proofread a part of my first draft of this thesis, and to Pontus Stenetorp, who carefully read and gave comments on my draft of the COLING paper.

I am also grateful to my lab mates, especially Han Dan and Sho Hoshino, as well as intern students to NII: Ying Xu, who always cheered me up, encouraged me to come to the lab early in every morning, and also helped me a lot in Beijing during ACL conference; and Le Quang Thang, whom I really enjoyed mentoring in our supervised parsing works.

During my research life in NII, the secretaries in Miyao lab, Keiko Kigoshi and Yuki Amano, not only greatly help in many office works, but also are a few connection to the outside of the

research. I was able to continue research in NII thanks to their kind considerations.

Many thanks to my colleagues and friends at Macquarie University, especially to Dat Quoc Nguyen, Kinzang Chhogyal, and Anish Kumar. I have enjoyed daily lunch and pub nights with you, with which I was able to stay sane without a lot of stress in my first long-term stay abroad. I am also indebted to John Pate, who helps me a lot at the beginning of my study on unsupervised grammar induction.

Finally, I would like to thank my parents, my sister, and my brother. Thank you for your infinite supports through my life, and thank you for making me who I am.

Contents

| | |
|---|------------|
| Abstract | iii |
| Acknowledgments | v |
| 1 Introduction | 1 |
| 1.1 Tasks and Motivations | 3 |
| 1.1.1 Examining language universality of center-embedding avoidance | 3 |
| 1.1.2 Unsupervised grammar induction | 3 |
| 1.2 What this thesis is not about | 5 |
| 1.3 Contributions | 5 |
| 1.4 Organization of the thesis | 6 |
| 2 Background | 7 |
| 2.1 Syntax Representation | 7 |
| 2.1.1 Context-free grammars | 7 |
| 2.1.2 Constituency | 9 |
| 2.1.3 Dependency grammars | 9 |
| 2.1.4 CFGs for dependency grammars and spurious ambiguity | 10 |
| 2.1.5 Projectivity | 11 |
| 2.2 Left-corner Parsing | 13 |
| 2.2.1 Center-embedding | 13 |
| 2.2.2 Left-corner parsing strategy | 16 |
| 2.2.3 Push-down automata | 17 |
| 2.2.4 Properties of the left-corner PDA | 19 |
| 2.2.5 Another variant | 21 |
| 2.2.6 Psycholinguistic motivation and limitation | 24 |
| 2.3 Learning Dependency Grammars | 28 |
| 2.3.1 Probabilistic context-free grammars | 28 |
| 2.3.2 CKY Algorithm | 29 |
| 2.3.3 Learning parameters with EM algorithm | 30 |
| 2.3.4 When the algorithms work? | 33 |
| 2.3.5 Split bilexical grammars | 34 |
| 2.3.6 Dependency model with valence | 39 |

| | | |
|----------|---|-----------|
| 2.3.7 | Log-linear parameterization | 40 |
| 2.4 | Previous Approaches in Unsupervised Grammar Induction | 41 |
| 2.4.1 | Task setting | 41 |
| 2.4.2 | Constituent structure induction | 42 |
| 2.4.3 | Dependency grammar induction | 43 |
| 2.4.4 | Other approaches | 46 |
| 2.4.5 | Summary | 47 |
| 3 | Multilingual Dependency Corpora | 49 |
| 3.1 | Heads in Dependency Grammars | 50 |
| 3.2 | CoNLL Shared Tasks Dataset | 52 |
| 3.3 | Universal Dependencies | 54 |
| 3.4 | Google Universal Treebanks | 55 |
| 4 | Left-corner Transition-based Dependency Parsing | 56 |
| 4.1 | Notations | 58 |
| 4.2 | Stack Depth of Existing Transition Systems | 59 |
| 4.2.1 | Arc-standard | 59 |
| 4.2.2 | Arc-eager | 60 |
| 4.2.3 | Other systems | 61 |
| 4.3 | Left-corner Dependency Parsing | 62 |
| 4.3.1 | Dummy node | 62 |
| 4.3.2 | Transition system | 62 |
| 4.3.3 | Oracle and spurious ambiguity | 66 |
| 4.3.4 | Stack depth of the transition system | 69 |
| 4.4 | Empirical Stack Depth Analysis | 70 |
| 4.4.1 | Settings | 71 |
| 4.4.2 | Stack depth for general sentences | 71 |
| 4.4.3 | Comparing with randomized sentences | 72 |
| 4.4.4 | Token-level and sentence-level coverage results | 72 |
| 4.4.5 | Results on UD | 76 |
| 4.4.6 | Relaxing the definition of center-embedding | 76 |
| 4.5 | Parsing Experiment | 78 |
| 4.5.1 | Feature | 80 |
| 4.5.2 | Settings | 81 |
| 4.5.3 | Results on the English Development Set | 82 |
| 4.5.4 | Result on CoNLL dataset | 84 |
| 4.5.5 | Result on UD | 87 |
| 4.6 | Discussion and Related Work | 91 |

| | | |
|----------|---|------------|
| 5 | Grammar Induction with Structural Constraints | 93 |
| 5.1 | Approach Overview | 94 |
| 5.1.1 | Structure-Constrained Models | 94 |
| 5.1.2 | Learning Structure-Constrained Models | 95 |
| 5.2 | Simulating split-bilexical grammars with a left-corner strategy | 96 |
| 5.2.1 | Handling of dummy nodes | 97 |
| 5.2.2 | Head-outward and head-inward | 97 |
| 5.2.3 | Algorithm | 98 |
| 5.2.4 | Spurious ambiguity and stack depth | 103 |
| 5.2.5 | Relaxing the definition of center-embedding | 103 |
| 5.3 | Experimental Setup | 104 |
| 5.3.1 | Datasets | 104 |
| 5.3.2 | Baseline model | 105 |
| 5.3.3 | Evaluation | 106 |
| 5.3.4 | Parameter-based Constraints | 107 |
| 5.3.5 | Structural Constraints | 109 |
| 5.3.6 | Other Settings | 110 |
| 5.4 | Empirical Analysis | 110 |
| 5.4.1 | Universal Dependencies | 110 |
| 5.4.2 | Qualitative analysis | 113 |
| 5.4.3 | Google Universal Dependency Treebanks | 119 |
| 5.5 | Discussion | 120 |
| 5.6 | Conclusion | 121 |
| 6 | Conclusions | 122 |
| A | Analysis of Left-corner PDA | 124 |
| B | Part-of-speech tagset in Universal Dependencies | 128 |

List of Tables

| | | |
|-----|---|-----|
| 3.1 | Overview of CoNLL dataset (mix of training and test sets). Punc. is the ratio of punctuation tokens in a whole corpus. Av. len. is the average length of a sentence. . | 53 |
| 3.2 | Overview of UD dataset (mix of train/dev/test sets). Punc. is the ratio of punctuation tokens in a whole corpus. Av. len. is the average length of a sentence. | 55 |
| 4.1 | Order of required stack depth for each structure for each transition system. $O(1 \sim n)$ means that it recognizes a subset of structures within a constant stack depth but demands linear stack depth for the other structures. | 61 |
| 4.2 | Token-level and sentence-level coverage results of left-corner oracles with depth_{re} . Here, the right-hand numbers in each column are calculated from corpora that exclude all punctuation, e.g., 92% of tokens in Arabic are covered within a stack depth ≤ 3 , while the number increases to 94.1 when punctuation is removed. Further, 57.6% of sentences (61.6% without punctuation) can be parsed within a maximum depth_{re} of three, i.e., the maximum degree of center-embedding is at most two in 57.6% of sentences. Av. len. indicates the average number of words in a sentence. . | 75 |
| 4.3 | Feature templates used in both full and restricted feature sets, with t representing POS tag and w indicating the word form, e.g., $s_0.l.t$ refers to the POS tag of the leftmost child of s_0 . \circ means concatenation. | 81 |
| 4.4 | Additional feature templates only used in the full feature model. | 81 |
| 4.5 | Parsing results on CoNLL X and 2007 test sets with no stack depth bound (unlabeled attachment scores). | 86 |
| 5.1 | Statistics on UD15 (after stripping off punctuations). Av. len. is the average length. Test ratio is the token ratio of the test set. | 105 |
| 5.2 | Statistics on Google trebanks (maximum length = 10). | 106 |
| 5.3 | Accuracy comparison on UD15 for selected configurations including harmonic initialization (Harmonic). Unif. is a baseline model without structural constraints. C is the allowed constituent length when the maximum stack depth is one. β_{len} is strength of the length bias. | 112 |
| 5.4 | Accuracy comparison on UD15 for selected configurations with the hard constraints on possible root POS tags. | 113 |
| 5.5 | Ratio of function words in the training corpora of UD (sentences of length 15 or less). | 119 |

| | | |
|-----|--|-----|
| 5.6 | Attachment scores on Google universal treebanks (up to length 10). All proposed models are trained with the verb-otherwise-noun constraint. Naseem10 = the model with manually crafted syntactic rules between POS tags (Naseem et al., 2010); Grave15 = also relies on the syntactic rules but is trained discriminatively (Grave and Elhadad, 2015). | 120 |
|-----|--|-----|

List of Figures

| | | |
|------|---|----|
| 2.1 | A set of rules in a CFG in which $N = \{S, NP, VP, VBD, DT, NN\}$, $\Sigma = \{\text{Mary, met, the, senator}\}$, and $S = S$ (the start symbol). | 8 |
| 2.2 | A parse tree with the CFG in Figure 2.1. | 8 |
| 2.3 | Example of labelled projective dependency tree. | 9 |
| 2.4 | Example of unlabelled projective dependency tree. | 9 |
| 2.5 | A set of template rules for converting dependency grammars into CFGs. a and b are lexical tokens (words) in the input sentence. $X[a]$ is a nonterminal symbol indicating the head of the corresponding span is a . | 10 |
| 2.6 | A CFG parse that corresponds to the dependency tree in Figure 2.4. | 11 |
| 2.7 | Another CFG parse that corresponds to the dependency tree in Figure 2.4. | 11 |
| 2.8 | Example of non-projective dependency tree. | 12 |
| 2.9 | The result of pseudo projectivization to the tree in Figure 2.8. | 12 |
| 2.10 | A parse involves center-embedding if the pattern in (a) is found in it. (b) and (c) are the minimal patterns with degree one and two respectively. (d) is the symmetry of (b) but we regard this as not center-embedding. | 15 |
| 2.11 | (a)–(c) Three kinds of branching structures with numbers on symbols and arcs showing the order of recognition with a left-corner strategy. (d) a partial parse of (c) using a left-corner strategy just after reading symbol b , with gray edges and symbols showing elements not yet recognized; The number of connected subtrees here is 2. | 16 |
| 2.12 | The set of transitions in a push-down automaton that parses a CFG (N, Σ, P, S) with the left-corner strategy. $a \in \Sigma$; $A, B, C, D \in N$. The initial stack symbol q_{init} is the start symbol of the CFG S , while the final stack symbol q_{final} is an empty stack symbol ε . | 18 |
| 2.13 | Graphical representations of inferences rules of PREDICTION and COMPOSITION defined in Figure 2.12. An underlined symbol indicates that the symbol is predicted top-down. | 18 |
| 2.14 | An example of parsing process by the left-corner PDA to recover the parse in Figure 2.10(b) given an input sentence $a b c d$. It is step 4 that occurs stack depth two after a reduce transition. | 19 |
| 2.15 | A CFG that is parsed with the process in Figure 2.14. | 19 |
| 2.16 | A set of transitions in another variant of the left-corner PDA appeared in Resnik (1992). $a \in \Sigma$; $A, B, C, D \in N$. Differently from the PDA in Figure 2.12, the initial stack symbol q_{init} is S while q_{final} is an empty stack symbol ε . | 21 |

| | | |
|------|--|----|
| 2.17 | Stack symbols of the left-corner PDA of Figure 2.16. Both trees correspond to symbol $A-B$ where A is the current goal while B is the recognized nonterminal. Note that A may be a right descendant of another nonterminal (e.g., X), which dominates a larger subtree. | 21 |
| 2.18 | Parsing process of the PDA in Figure 2.16 to recover the parse in Figure 2.10(b) given the CFG in Figure 2.15 and an input sentence $a b c d$. The stack depth keeps one in every step after a shift transition. | 22 |
| 2.19 | Parsing process of the PDA in Figure 2.16 to recover the parse in Figure 2.10(d). The stack depth after a shift transition increases at step 3. | 23 |
| 2.20 | The parse of the sentence (2a). | 25 |
| 2.21 | The parse of the sentence (5). | 26 |
| 2.22 | Inference rules of the CKY algorithm. TERMINAL rules correspond to the terminal expansion in line 8 of the Algorithm 1; BINARY rules correspond to the one in line 10. Each rule specifies how an analysis of a larger span (below \rightarrow) is derived from the analyses of smaller spans (above \rightarrow) provided that the input and grammar satisfy the side conditions in the right of \rightarrow | 31 |
| 2.23 | Example of a projective dependency tree generated by a SBG. $\$$ is always placed at the end of a sentence, which has only one dependent in the left direction. | 34 |
| 2.24 | Binary inference rules in the naive CFG conversion. F means the state is a final state in that direction. Both left and right consequent items (below \rightarrow) have the same item but from different derivations, suggesting 1) the weighted grammar is not a PCFG; and 2) there is the spurious ambiguity. | 36 |
| 2.25 | An algorithm for parsing SBGs in $O(n^3)$ given a length n sentence. The $n + 1$ -th token is a dummy root token $\$$, which only has one left dependent (sentence root). i, j, h , and h' are index of a token in the given sentence while q, r , and F are states. L_h and R_h are left and right FSA of the h -th token in the sentence. Each item as well as a statement about a state (e.g., $r \in \text{final}(L_p)$) has a weight and the weight of a consequent item (below \rightarrow) is obtained by the product of the weights of its antecedent items (above \rightarrow). | 37 |
| 2.26 | Mappings between FSA transitions of SBGs and the weights to achieve DMV. θ_s and θ_a are parameters of DMV described in the body. The right cases (e.g., $q_0 \in \text{init}(R_a)$) are omitted but defined similarly. h and d are both word types, not indexes in a sentence (contrary to Figure 2.25). | 39 |
| 2.27 | An example of bootstrapping process for assigning category candidates in CCG induction borrowed from Bisk and Hockenmaier (2013). DT, NNS, VBD, and RB are POS tags. Bold categories are the initial seed knowledge, which is expanded by allowing the neighbor token to be a modifier. | 47 |
| 3.1 | Each dataset that we use employs the different kind of annotation style. Bold arcs are ones that do not exist in the CoNLL style tree (a). | 50 |
| 3.2 | A dependency tree in the Japanese UD. NOUN, ADV, VERB, and ADP are assigned POS tags. | 51 |
| 3.3 | Four styles of annotation for coordination. | 51 |

| | | |
|------|---|----|
| 4.1 | Conversions from dependency trees into CFG parses; (a) can be uniquely converted to (b), while (c) can be converted to both (d) and (e). | 59 |
| 4.2 | (a)-(c) Right-branching dependency trees for three words and (d) the corresponding CFG parse. | 60 |
| 4.3 | Example configuration of a left-corner transition system. | 63 |
| 4.4 | Actions of the left-corner transition system including two shift operations (top) and reduce operations (bottom). | 63 |
| 4.5 | Correspondences of reduce actions between dependency and CFG. We only show minimal example subtrees for simplicity. However, a can have an arbitrary number of children, so can b or x , provided x is on a right spine and has no right children. . | 64 |
| 4.6 | An example parsing process of the left-corner transition system. | 65 |
| 4.7 | Implicit binarization process of the oracle described in the body. | 68 |
| 4.8 | Center-embedded dependency trees and zig-zag patterns observed in the implicit CFG parses: (a)–(b) depth one, (c)–(d) depth two, (e) CFG parse for (a) and (b), and (f) CFG parse for (c) and (d). | 70 |
| 4.9 | Crosslinguistic comparison of the cumulative frequencies of stack depth during oracle transitions. | 73 |
| 4.10 | Stack depth results in corpora with punctuation removed; the dashed lines show results on randomly reordered sentences. | 74 |
| 4.11 | Stack depth results in UD. | 77 |
| 4.12 | Following Definition 2.2, this tree is recognized as singly center-embedded while is not center-embedded if “the senator” is replaced by one word. Bold arcs are the cause of center-embedding (zig-zag pattern). | 78 |
| 4.13 | Stack depth results in UD with a left-corner system (depth_{re}) when the definition of center-embedding is relaxed. The parenthesized numbers indicate the size of allowed constituents at the bottom of embedding. For example (2) next to 2 indicates we allow $\text{depth} = 3$ if the size of subtree on the top of the stack is 1 or 2. Len. is the maximum sentence length. | 79 |
| 4.14 | (Left) Elementary features extracted from an incomplete and complete node, and (Right) how feature extraction is changed depending on whether the next step is shift or reduce. | 80 |
| 4.15 | Accuracy vs. stack depth bound at decoding for several beam sizes (b). | 83 |
| 4.16 | Accuracy vs. beam size for each system on the English Penn Treebank development set. Left-corner (full) is the model with the full feature set, while Left-corner (limited) is the model with the limited feature set. | 84 |
| 4.17 | Accuracy vs. stack depth bound in CoNLL dataset. | 85 |
| 4.18 | (a)-(b) Example of a parse error by the left-corner parser that may be saved with external syntactic knowledge (limited features and beam size 8). (c) Two corresponding configuration paths: the left path leads to (a) and the right path leads to (b). | 88 |
| 4.19 | Accuracy vs. stack depth bound in UD. | 89 |
| 4.20 | Accuracy vs. stack depth bound with left-corner parsers on UD with different maximum length of test sentences. | 90 |

| | | |
|-----|--|-----|
| 5.1 | Dummy nodes (x in (a) and (b)) in the transition system cannot be used in our transition system because with this method, we have to remember every child token of the dummy node to calculate attachment scores at the point when the dummy is filled with an actual token, which leads to an exponential complexity. We instead abstract trees in a different way as depicted in (c) by not abstracting the predicted node p but filling with the actual word (p points to some index in a sentence such that $j < p \leq n$). If $i = 1, j = 3$, this representation abstracts both tree forms of (a) and (b) with some fixed x (corresponding to p). | 98 |
| 5.2 | An algorithm for parsing SBGs with a left-corner strategy in $O(n^4)$ given a sentence of length n , except the composition rules which are summarized in Figure 5.3. The $n + 1$ -th token is a dummy root token $\$$, which only has one left dependent (sentence root). i, j, h, p are indices of tokens. The index of a head which is still collecting its dependents is decorated with a state (e.g., q). L_h and R_h are left and right FSAs of SBGs given a head index h , respectively; we reverse the process of L_h to start with $q \in \text{final}(L_h)$ and finish with $q \in \text{init}(L_h)$ (see the body). Each item is also decorated with depth d that corresponds to the stack depth incurred when building the corresponding tree with left-corner parsing. Since an item with larger depth is only required for composition rules, the depth is unchanged with the rules above, except SHIFT-*, which corresponds to SHIFT transition and can be instantiated with arbitrary depth. Note that ACCEPT is only applicable for an item with depth 1, which guarantees that the successful parsing process remains a single tree on the stack. Each item as well as a statement about a state (e.g., $r \in \text{final}(L_p)$) has a weight and the weight of a consequence item is obtained by the product of the weights of its antecedent items. | 99 |
| 5.3 | The composition rules that are not listed in Figure 5.2. LEFTCOMP is divided into two cases, LEFTCOMP-L-* and LEFTCOMP-R-* depending on the position of the dummy (predicted) node on the left antecedent item (corresponding to the second top element on the stack). They are further divided into two processes, 1 and 2 for achieving head-splitting. b is an additional annotation on an intermediate item for correct depth computation in LEFTCOMP. | 100 |
| 5.4 | We decompose the LEFTCOMP action defined for the transition system into two phases, LEFTCOMP-L-1 and LEFTCOMP-L-2, each of which collects only left or right half constituent of a subtree on the top of the stack. A number above each stack element is the stack depth decorated on the corresponding chart item. | 101 |
| 5.5 | Attachment accuracies on UD15 with the function word constraint and structural constraints. The numbers in parentheses are the maximum length of a constituent allowed to be embedded. For example (3) means a part of center-embedding of depth two, in which the length of embedded constituent ≤ 3 , is allowed. | 111 |
| 5.6 | Comparison of output parses of several models on a sentence in English UD. The outputs of $C = 2$ and $C = 3$ do not change with the root POS constraint, while the output of $\beta_{len} = 0.1$ changes to the same one of the uniform model with the root POS constraint. Colored arcs indicate the wrong predictions. Note surface forms are not observed by the models (only POS tags are). | 115 |

| | | |
|-----|--|-----|
| 5.7 | Another comparison between outputs of the uniform model and $C = 3$ in English UD. We also show $\beta_{len} = 0.1$ for comparison. Although the score difference is small (see Table 5.3), the types of errors are different. In particular the most of parse errors by $C = 3$ are at local attachments (first-order). For example it consistently recognizes a noun is a head of a verb, and a noun is a sentence root. Note an error on “power \rightarrow purposes” is an example of PP attachment errors, which may not be solved under the current problem setting receiving only a POS tag sequence. . . . | 116 |
| A.1 | Three types of realizations of Eq. A.1. Dashed edges may consist of more than one edge (see Figure A.2 for example) while dotted edges may not exist (or consist of more than one edge). (a) E is a right child of C_{m_e} and thus the degree of center-embedding is m_e . (b) E is a left child of B_{m_e} (i.e., $C_{m_e} = E$) and the degree is $m_e - 1$; when $m = 1$, $C_1 = E$ and thus no center-embedding occurs. (c) $B_{m_e} = C_{m_e} = E$; note this happens only when $m_e = 1$ (see body). | 125 |
| A.2 | (a) Example of realization of a path between A and B_1 in Figure A.1. (b) The one between B_1 and C_1 | 126 |

Chapter 1

Introduction

Explaining the syntactic variation and universals including the constraints on that variation across languages in the world is essential both from a theoretical and practical point of view. It is in fact one of the main goals in linguistics (Greenberg, 1963; Dryer, 1992; Evans and Levinson, 2009). In computational linguistics, these kinds of syntactic regularities and constraints could be utilized as prior knowledge about grammars, which would be valuable for improving the performance of various syntax-oriented systems such as parsers or grammar induction systems, e.g., by being encoded as *features* in a system (Collins, 1999; McDonald et al., 2005). This thesis is about such syntactic universals.

Our goal in this thesis is to identify a good syntactic constraint that fits well to the natural language sentences and thus could be exploited to improve the performance of syntax-oriented systems such as parsers. For this end, we pick up a well known linguistic phenomenon that might be universal across languages, empirically examine its language universality across diverse languages using cross-linguistic datasets, and present computational experiments to demonstrate its utility in a real application. Along with this, we also define several computational algorithms that efficiently exploit the constraint during sentence processing. For an application, we show that our constraint will help in the task of unsupervised syntactic parsing, or grammar induction where the goal is to find salient syntactic patterns without explicit supervision about grammars.

In linguistics, one pervasive hypothesis about the origin of such syntactic constraints is that they come from the limitations on the human cognitive mechanism and pressures associated with language acquisition and use (Jaeger and Tily, 2011; Fedzechkina et al., 2012). In other words, since the language is a tool for communication, it is natural to assume that its shape has been formed to increase the daily communication efficiency or the learnability for language learners. The underlying commonalities in diverse languages are then understood as the outcome of such pressures that every language user might naturally suffer from. Our focused constraint in this thesis also has its origin in the restriction of the human ability of comprehension observed in several psycholinguistic experiments, which we introduce next.

Center-embedding It is well known in the psycholinguistic literature that a nested, or center-embedded structure is particularly difficult for comprehension:

- (1) # The reporter [who the senator [who Mary met attacked] ignored] the president.

This sentence is called center-embedding by its syntactic construction indicated with brackets. This observation will be the starting point of the current study. The difficulty of center-embedded structures has been testified across a number of languages including English (Gibson, 2000; Chen et al., 2005) and Japanese (Nakatani and Gibson, 2010). Compared to these behavioral studies, the current study aims to characterize the phenomenon of center-embedding from *computational* and quantitative perspectives. For instance, one significance of the current study is to show that center-embedding is in fact a rarely observed syntactic phenomenon across a variety of languages. We verify this fact using syntactically annotated corpora, i.e., treebanks of more than 20 languages.

Left-corner Another important concept in this thesis is *left-corner* parsing. A left-corner parser parses an input sentence on a *stack*; the distinguished property of it is that its stack depth increases only when generating, or accepting center-embedded structures. These formal properties of left-corner parsers were studied more than 20 years ago (Abney and Johnson, 1991; Resnik, 1992) although until now there exists little study concerning its empirical behaviors as well as its potential for a device to exploit syntactic regularities of languages as we investigate in this thesis. One previous attempt for utilizing a left-corner parser in a practical application is Johnson’s linear-time tabular parsing by approximating the state space of a parser by a finite state machine. However, this trial was not successful (Johnson, 1998a).¹

Dependency Our empirical examinations listed above will be based on the syntactic representation called *dependency* structures. In computational linguistics, constituent structures have long played a central role as a representation of natural language syntax (Stolcke, 1995; Collins, 1997; Johnson, 1998b; Klein and Manning, 2003) although this situation has been changed and the recent trend in the parsing community has favored dependency-based representations, which are conceptually simpler and thus often lead to more efficient algorithms (Nivre, 2003; Yamada and Matsumoto, 2003; McDonald et al., 2005; Goldberg and Nivre, 2013). Another important reason for us to focus on this representation is that its *unsupervised* induction is more tractable than the constituent representation, such as phrase-structure grammars. In fact, significant research on unsupervised parsing has been done in this decade though much of it assumes dependency trees as the underlying structure (Klein and Manning, 2004; Smith and Eisner, 2006; Berg-Kirkpatrick et al., 2010; Mareček and Žabokrtský, 2012; Spitkovsky et al., 2013). We discuss this computational issue more in the next chapter (Section 2.4).

The last, and perhaps the most essential advantage of a dependency representation is its cross-linguistic suitability. For studying the empirical behavior of some system across a variety of languages, the resources for those languages are essential. Compared to constituent structures, dependency annotations are available in many corpora covering more than 20 languages across diverse language families. Each treebank typically contains thousands of sentences with manually parsed

¹The idea is that since a left-corner parser can recognize most of (English) sentences within a limited stack depth bound, e.g., 3, the number of possible stack configurations will be constant and we may construct a finite state machine for a given context-free grammar. However in practice, the grammar constant for this algorithm gets much larger, leading to $O(n^3)$ asymptotic runtime, the same as the ordinary parsing method, e.g., CKY.

syntactic trees. We use such large datasets to examine our hypotheses about universal properties of languages. Though the concepts introduced above, center-embedding and left-corner parsing, are both originally defined on constituent structures, we describe in this thesis a method by which they can be extended to dependency structures via a close connection between two representations.

1.1 Tasks and Motivations

More specifically, the tasks we tackle in this thesis can be divided into the following two categories, each of which is based on specific motivations.

1.1.1 Examining language universality of center-embedding avoidance

We first examine the hypothesis that center-embedding is a language phenomenon that every language user tries to avoid *regardless* of language. The quantitative study for this question across diverse languages has not yet been performed. Two motivations exist for this analysis: One is rather scientific: we examine the explanatory power of center-embedding avoidance as a universal grammatical constraint. This is ambitious though we put more weight on the second, rather practical motivation: the possibility that avoidance of center-embedding is a good syntactic bias to restrict the space of possible tree structures of natural language sentences. These analyses are the main topic of Chapter 4.

1.1.2 Unsupervised grammar induction

We then consider applying the constraint with center-embedding into the application of *unsupervised grammar induction*. In this task, the input to the algorithm is a collection of sentences, from which the model tries to extract the salient patterns as a grammar. This setting contrasts with the more familiar *supervised* parsing task in which typically some machine learning algorithm learns the mapping from a sentence to the syntactic tree based on the training examples, i.e., sentences paired with their corresponding parse trees. In the unsupervised setting, our goal is to obtain a model that can parse a sentence without access to the correct trees for training sentences. This is a particularly hard problem though we expect the universal syntactic constraint may help in improving the performance since it can effectively restrict the possible search space for the model.

Motivations A typical reason to tackle this task is a purely engineering one: Although the number of languages that we can access to the resource (i.e., treebank) increases, there are still so many languages in the world for which little to no resources are available since the creation of a new treebank from scratch is still very hard and time consuming. Unsupervised learning of grammars would be helpful for this situation, as it provides a cheap solution without requiring the manual efforts of linguistic experts. A more realistic setting might be to use the output of an unsupervised system as the initial annotation, which could then be corrected by experts later. In short, a better unsupervised system can reduce the effort of experts in preparing new treebanks. This motivation can be held in any efforts of unsupervised grammar induction.

However, as we do in this thesis, the grammar induction with particular syntactic biases or constraints would also be appealing for the following reasons as well:

- We can regard this task as a typical example of more broad problems of learning syntax without explicit supervision. An example of such problem is a grounding task, in which the learner induces the model of (intermediate) tree structures that bridge an input sentence and its semantics, which may be represented in one of several different forms, depending on the task and corpus, e.g., the logical expression (Zettlemoyer and Collins, 2005; Kwiatkowski et al., 2010) and the answer to the given question (Liang et al., 2011; Berant et al., 2013; Kwiatkowski et al., 2013; Kushman et al., 2014). In these tasks, though some implicit supervision is provided, the search space is typically still very large. Obtaining a positive result for the current unsupervised learning, we argue, would present an important starting point for extending the current idea into such related grammar induction tasks. What type of supervision we should give for those tasks is also still an open problem; one possibility is that a good *prior* for general natural language syntax, as we investigate here, would reduce the amount of supervision necessary for successful learning. Finally, we claim that although the current study focuses on inducing dependency structures, the presented idea, avoiding center-embedding during learning, is general enough and not necessarily restricted to the dependency induction tasks. The main reason why we focus on dependency structures is rather computational (see Section 2.4), but it may not hold in the grounded learning tasks in the previous works cited above. Moreover, recently more sophisticated grammars such as combinatory categorical grammars (CCGs) are shown to be learnable when appropriate light supervision is given as seed knowledge (Bisk and Hockenmaier, 2013; Bisk and Hockenmaier, 2015; Garrette et al., 2015). We thus believe that the lesson from the current study will also shed light on those related learning tasks that do not assume dependency trees as the underlying structures.
- The final motivation is in the relevance to understanding of child language acquisition. Computational modeling of the language acquisition process, in particular using probabilistic models, has gained much attention in recent years (Goldwater et al., 2009; Kwiatkowski et al., 2012; Johnson et al., 2012; Doyle and Levy, 2013). Although many of those works cited above focus on modeling of relatively early acquisition problems, e.g., word segmentation from phonological inputs, some initial studies regarding acquisition mechanism of grammar also exist (Pate and Goldwater, 2013).

We argue here that our central motivation is *not* to get insights into the language acquisition mechanism although the structural constraint that we consider in this thesis (i.e., center-embedding) originally comes from observation of human sentence processing. This is because our problem setting is far from the real language acquisition scenario that a child may undergo. There exist many discrepancies between them; the most problematic one is found in the input to the learning algorithm. For resource reasons, the input sentences to our learning algorithm are largely written texts for adults, e.g, newswires, novels, and blogs. This contrasts with the relevant studies cited above on word segmentation in which the input for training is phonological forms of child directed speech, which is, however, available in only a few languages such as English. This poses a problem since our interest in this thesis is the

language universality of the constraint, which needs many language treebanks to be evaluated. Another limitation of the current approach is that every model in this thesis assumes the part-of-speech (POS) of words in a sentence as its input rather than the surface form. This simplification makes the problem much simpler and is employed in most previous studies (Klein and Manning, 2004; Smith and Eisner, 2006; Berg-Kirkpatrick et al., 2010; Bisk and Hockenmaier, 2013; Grave and Elhadad, 2015), but it is of course an unrealistic assumption about inputs that children receive.

Our main claim in this direction is that the success of the current approach would lead to the further study about the connection between the child language acquisition and computational modeling of the acquisition process. We leave the remaining discussion about this topic in the conclusion of this thesis.

1.2 What this thesis is not about

This thesis is not about psycholinguistic study, i.e., we do not attempt to reveal the mechanism of human sentence processing. Our main purpose in referring to the literature in psycholinguistics is to get insights for the syntactic patterns that every language might share to some extent and thus could be exploited from a system of computational linguistics. We would be pleased if our findings about the universal constraint affect the thinking of psycholinguists but this is not the main goal of the current thesis.

1.3 Contributions

Our contributions can be divided into the following four parts. The first two are our conceptual, or algorithmic contributions while the latter two are the contributions of our empirical study.

Left-corner dependency parsing algorithm We show how the idea of left-corner parsing, which was originally developed for recognizing constituent structures, can be extended to dependency structures. We formalize this algorithm in the framework of transition-based parsing, a similar device to the pushdown automata often used to describe the parsing algorithm for dependency structures. The resulting algorithm has the property that its stack depth captures the degree of center-embedding of the recognizing structure.

Efficient dynamic programming We extend this algorithm into the tabular method, i.e., chart parsing, which is necessary to combine the ideas of left-corner parsing and unsupervised grammar induction. In particular, we describe how the idea of head splitting (Eisner and Satta, 1999; Eisner, 2000), a technique to reduce the time complexity of chart-based dependency parsing, can be applied in the current setting.

Evidence on the universality of center-embedding avoidance We show that center-embedding is a rare construction across languages using treebanks of more than 20 languages. Such large-scale investigation has not been performed before in the literature. Our experiment is composed of two types of complementary analyses: a static, counting-based analysis of treebanks

and a supervised parsing experiment to see the effect of the constraint when some amount of parse errors occurs.

Unsupervised parsing experiments with structural constraints We finally show that our constraint does improve the performance of unsupervised induction of dependency grammars in many languages.

1.4 Organization of the thesis

The following chapters are organized as follows.

- In Chapter 2, we summarize the backgrounds necessary to understand the following chapters of the thesis including several syntactic representations, the EM algorithm for acquiring grammars, and left-corner parsing.
- In Chapter 3, we summarize the multilingual corpora we use in our experiments in the following chapters.
- Chapter 4 covers the topics of the first and third contributions in the previous section. We first define a tool, i.e., a left-corner parsing algorithm for dependency structures, for our corpus analysis in the remainder of the chapter.
- Chapter 5 covers the remaining, second and fourth contributions in the previous section. Our experiments on unsupervised parsing require the formulation of the EM algorithm, which relies on chart parsing for calculating sufficient statistics. We thus first develop a new dynamic programming algorithm and then apply it to the unsupervised learning task.
- Finally, in Chapter 6, we summarize the results obtained in this research and give directions for future studies.

Chapter 2

Background

The topics covered in this chapter can be largely divided into four parts. Section 2.1 defines several important concepts for representing syntax, such as constituency and dependency, which become the basis of all topics discussed in this thesis. We then discuss left-corner parsing and related issues in Section 2.2, such as the formal definition of center-embedding, which are in particular important to understand the contents in Chapter 4. The following two sections are more relevant to our application of unsupervised grammar induction discussed in Chapter 5. In Section 2.3, we describe the basis of learning probabilistic grammars, such as the EM algorithm. Finally in Section 2.4, we provide the thorough survey of the unsupervised grammar induction, and make clear our motivation and standpoint for this task.

2.1 Syntax Representation

This section introduces several representations to describe the natural language syntax appearing in this thesis, namely context-free grammars, constituency, and dependency grammars, and discuss the connection between them. Though our focused representation in this thesis is dependency, the concepts of context-free grammars and constituency are also essential for us. For example, context-free grammars provide the basis for probabilistic modeling of tree structures as well as parameter estimation for it; We discuss how our dependency-based model can be represented as an instance of context-free grammars in Section 2.3. The connection between constituency and dependency appears many times in this thesis. For instance, the concept of *center-embedding* (Section 2.2) is more naturally understood with constituency rather than with dependency.

This section is about syntax representation or grammars and we do not discuss *parsing* but to see how the analysis with a grammar looks like, we mention a *parse* or a parse tree, which is the result of parsing for an input string (sentence).

2.1.1 Context-free grammars

A context-free grammar (CFG) is a useful model to describe the hierarchical syntactic structure of an input string (sentence). Formally a CFG is a quadruple $G = (N, \Sigma, P, S)$ where N and Σ are disjoint finite set of symbols called nonterminal and terminal symbols respectively. Terminal

| | | |
|-----|---|---------|
| S | → | NP VP |
| VP | → | VBD NP |
| NP | → | DT NN |
| NP | → | Mary |
| VBD | → | met |
| DT | → | the |
| NN | → | senator |

Figure 2.1: A set of rules in a CFG in which $N = \{S, NP, VP, VBD, DT, NN\}$, $\Sigma = \{\text{Mary, met, the, senator}\}$, and $S = S$ (the start symbol).

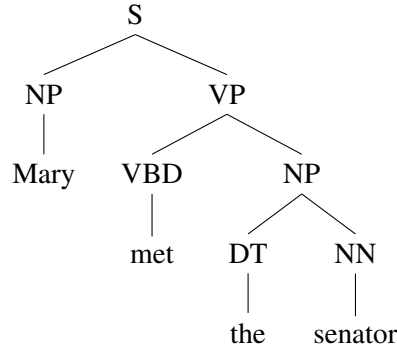


Figure 2.2: A parse tree with the CFG in Figure 2.1.

symbols are symbols that appear at leaf positions of a tree while nonterminal symbols appear at internal positions. $S \in N$ is the start symbol. P is the set of rules of the form $A \rightarrow \beta$ where $A \in N$ and $\beta \in (N \cup \Sigma)^*$.

Figure 2.1 shows an example of a CFG while Figure 2.2 shows an example of a parse with that CFG. On a parse tree *terminal nodes* refer to the nodes with terminal symbols (at leaf positions) while *nonterminal nodes* refer to other internal nodes with nonterminal symbols. *Preterminal nodes* are nodes that appear just above terminal nodes (e.g., VBD in Figure 2.2).

This model is useful because there is a well-known polynomial (cubic) time algorithm for parsing an input string with it, which also provides the basis for parameters estimation when we develop probabilistic models on CFGs (see Section 2.3.3).

Chomsky normal form A CFG is said to be in Chomsky normal form (CNF) if every rule in P has the form $A \rightarrow BC$ or $A \rightarrow a$ where $A, B, C \in N$ and $a \in \Sigma$; that is, every rule is a binary rule or a unary rule and a unary rule is only allowed on a preterminal node. The CFG in Figure 2.1 is in CNF. We often restrict our attention to CNF as it is closely related to projective dependency grammars, our focused representation described in Section 2.1.3.

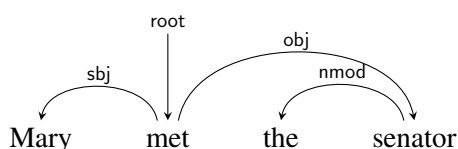


Figure 2.3: Example of labelled projective dependency tree.

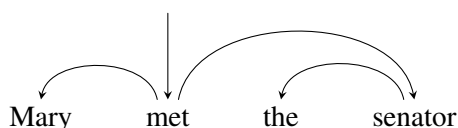


Figure 2.4: Example of unlabelled projective dependency tree.

2.1.2 Constituency

The parse in Figure 2.2 also describes the *constituent* structure of the sentence. Each constituent is a group of consecutive words that function as a single cohesive unit. In the case of tree in Figure 2.2, each constituent is a phrase spanned by some nonterminal symbol (e.g., “the senator” or “met the senator”).

We can see that the rules in Figure 2.1 define how a smaller constituents combine to form a larger constituent. This grammar is an example of *phrase-structure grammars*, in which each nonterminal symbol such as NP and VP describes the syntactic role of the constituent spanned by that nonterminal. For example, NP means the constituent is a noun phrase while VP means the one is a verb phrase. The phrase-structure grammar is often contrasted with dependency grammars, but we note that the concept of constituency is not restricted to phrase-structure grammars and plays an important role in dependency grammars as well, as we describe next.

2.1.3 Dependency grammars

Dependency grammars analyze the syntactic structure of a sentence as a directed tree of word-to-word dependencies. Each dependency is represented as a directed arc from a *head* to a *dependent*, which is argument or adjunct and is modifying the head syntactically or semantically. Figure 2.3 shows an example of an analysis with a dependency grammar. We call these directed trees *dependency trees*.

The question of what is the head is a matter of debate in linguistics. In many cases this decision is generally agreed but the analysis of certain cases is not settled, in particular those around function words (Zwicky, 1993). For example although “senator” is the head of the dependency between “the” and “senator” in Figure 2.3 some linguists argue “the” should be the head (Abney, 1987). We discuss this problem more in Chapter 3 where we describe the assumed linguistic theory in each treebank used in our experiments. See also Section 5.3.3 where we discuss that such discrepancies in head definitions cause a problem in evaluation for unsupervised systems (and our solution for that).

| Rewrite rule | Semantics |
|------------------------------|---|
| $S \rightarrow X[a]$ | Select a as the root word. |
| $X[a] \rightarrow X[a] X[b]$ | Select b as a right modifier of a . |
| $X[a] \rightarrow X[b] X[a]$ | Select b as a left modifier of a . |
| $X[a] \rightarrow a$ | Generate a terminal symbol. |

Figure 2.5: A set of template rules for converting dependency grammars into CFGs. a and b are lexical tokens (words) in the input sentence. $X[a]$ is a nonterminal symbol indicating the head of the corresponding span is a .

Labelled and unlabelled tree If each dependency arc in a dependency tree is annotated with a label describing the syntactic role between two words as in Figure 2.3, that tree is called a *labeled* dependency tree. For example the *sbj* label between “Mary” and “met” describes the subject-predicate relationship. A tree is called *unlabeled* if those labels are omitted, as in Figure 2.4.

In the remainder of this thesis, we only focus on *unlabeled* dependency trees although now most existing dependency-based treebanks provide labeled annotations of dependency trees. For our purpose, dependency labels do not play the essential role. For example, our analyses in Chapter 4 are based only on the tree shape of dependency trees, which can be discussed with unlabeled trees. In the task of unsupervised grammar induction, our goal is to induce the unlabeled dependency structures as we discuss in detail in Section 2.4.

Constituents in dependency trees The idea of constituency (Section 2.1.2) is not limited to phrase-structure grammars and we can identify the constituents in dependency trees as well. In dependency trees, a constituent is a phrase that comprises of a head and its descendants. For example, “met the senator” in Figure 2.4 is a constituent as it comprises of a head “met” and its descendants “the senator”. Constituents in dependency trees may be more directly understood by considering a CFG for dependency grammars and the parses with it, which we describe in the following.

2.1.4 CFGs for dependency grammars and spurious ambiguity

Figure 2.6 shows an example of a CFG parse, which corresponds to the dependency tree in Figure 2.4 but looks very much like the constituent structure in Figure 2.2. With this representation, it is very clear that *the senator* or *met the senator* is a constituent in the tree. We often rewrite an original dependency tree in this CFG form to represent the underlying constituents explicitly, in particular when discussing the extension of the concept of center-embedding and left-corner algorithm, which have originally assumed (phrase-structure-like) constituent structure, to dependency.

In this parse, every rewrite rule has one of the forms in Figure 2.5. Each rule specifies one dependency arc between a head and a dependent. For example, the rule $X[\text{senator}] \rightarrow X[\text{the}] X[\text{senator}]$ means that “senator” takes “the” as its left dependent.

Spurious ambiguity On the tree in Figure 2.4, we can identify “Mary met” is also a constituent, which is although not a constituent in the parses in Figure 2.6 and Figure 2.5. This divergence is

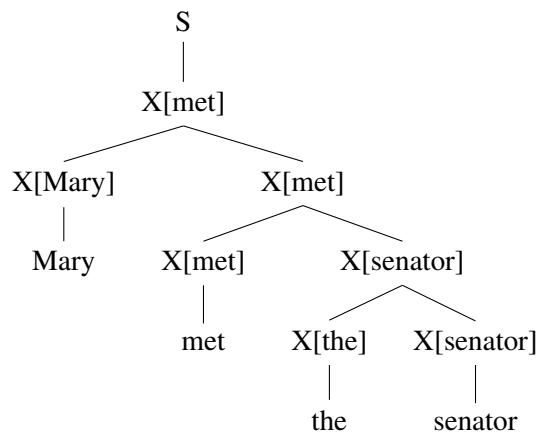


Figure 2.6: A CFG parse that corresponds to the dependency tree in Figure 2.4.

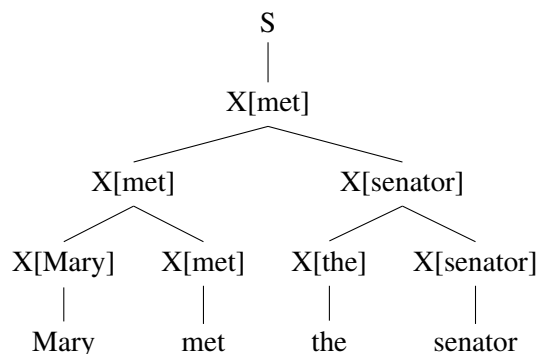


Figure 2.7: Another CFG parse that corresponds to the dependency tree in Figure 2.4.

related to the problem of *spurious ambiguity*, which indicates each dependency tree may correspond to more than one CFG parse. In fact, we can also build a CFG parse corresponding to Figure 2.4, in which contrary to Figure 2.2 the constituent of “Mary met” is explicitly represented with the nonterminal $X[\text{met}]$ dominating “Mary met”.

This ambiguity becomes the problem when we analyze the type of structure for a given dependency tree, e.g., whether a tree contains any center-embedded constructions. We will see the details and our solution for this problem later in Sections 4.3.3 and 4.3.4. Another related issue with this ambiguity is that it prevents us to use the EM algorithm for learning of the models built on this CFG, which we discuss in detail in Section 2.3.5.

2.1.5 Projectivity

A dependency tree is called *projective* if the tree does not contain any *crossing dependencies*. Every dependency tree appeared so far is projective. An example of non-projective tree is shown in Figure

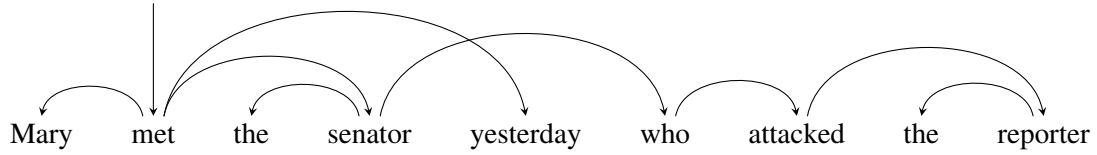


Figure 2.8: Example of non-projective dependency tree.

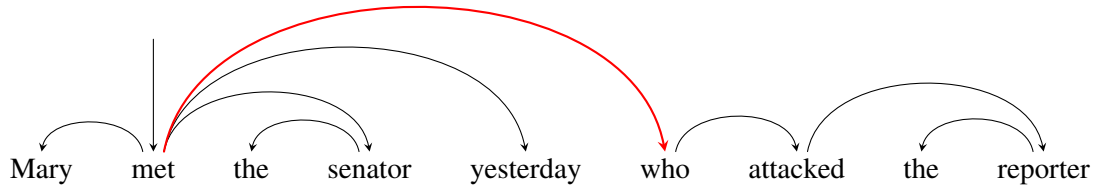


Figure 2.9: The result of pseudo projectivization to the tree in Figure 2.8.

2.8. Though we have not mentioned explicitly, the conversion method above can only handle projective dependency trees. If we allow non-projective structures in our analysis, then the model or the algorithm typically gets much more complex (McDonald and Satta, 2007; Gómez-Rodríguez et al., 2011; Kuhlmann, 2013; Pitler et al., 2013).¹

Non-projective constructions are known to be relatively rare cross-linguistically (Nivre et al., 2007a; Kuhlmann, 2013). Thus, along with the mathematical difficulty for handling them, often the dependency parsing algorithm is restricted to deal with only projective structures. For example, as we describe in Section 2.4, most existing systems of unsupervised dependency induction restrict their attention only on projective structures. Note that existing treebanks contain non-projective structures in varying degree so the convention is to restrict the model to generate only projective trees and to evaluate its quality against the (possibly) non-projective gold trees. We follow this convention in our experiments in Chapter 5 and generally focus only on projective dependency trees in other chapters as well, if not mentioned explicitly.

Pseudo-projectivity There is a known technique called pseudo-projectivization (Nivre and Nilsson, 2005), which converts any non-projective dependency trees into some projective trees with minimal modifications. The tree in Figure 2.9 shows the result of this procedure into the non-projective tree in Figure 2.8.² We perform this conversion on every tree when we analyze the

¹ The maximum spanning tree (MST) algorithm (McDonald et al., 2005) enables non-projective parsing in time complexity $O(n^2)$, which is more efficient than the ordinary CKY-based algorithm (Eisner and Satta, 1999) though the model form (i.e., features or conditioning contexts) is restricted to be quite simple.

² In the original formalization, pseudo-projectivization also performs label conversions. That is, the label on a (modified) dependency arc is changed for memorizing the performed operations; With this memorization, the converted tree does not lose the information. Nivre and Nilsson (2005) show that non-projective dependency parsing is possible with this conversion and parsing algorithms that assume projectivity, by training and decoding with the converted forms and recovering the non-projective trees from the labeled (projective) outputs. Since our focus is basically only unlabeled

structural properties of dependency trees in existing corpora in Chapter 4. See Section 4.4.1 for details.

2.2 Left-corner Parsing

In this section we describe left-corner parsing and summarize related issues, e.g., its relevance to psycholinguistic studies. Previous studies on left-corner parsing have focused only on a (probabilistic) CFG; We will extend the discussion in this section for dependency grammars in later chapters. In Chapter 4, we extend the idea into transition-based dependency parsing while in Chapter 5, we further extend the algorithm with efficient tabulation (dynamic programming).

A somewhat confusing fact about left-corner parsing is that there exist two variants of very different algorithms, called arc-standard and arc-eager algorithms. The arc-standard left-corner parsing has been appeared first in the programming language literature (Rosenkrantz and Lewis, 1970; Aho and Ullman, 1972) and later extended for natural language parsing for improving efficiency (Nederhof, 1993) or expanding contexts captured by the model (Manning and Carpenter, 2000; Henderson, 2004). In the following we do *not* discuss these arc-standard algorithms, and only focus on the arc-eager algorithm, which has its origin in psycholinguistics (Johnson-Laird, 1983; Abney and Johnson, 1991)³ rather than in computer science.

Left-corner parsing is closely relevant to the notion of center-embedding, a kind of branching pattern, which we characterize formally in Section 2.2.1. We then introduce the idea of left-corner parsing through a parsing strategy in Section 2.2.2 for getting intuition into parser behavior. During Sections 2.2.3 – 2.2.5, we discuss the push-down automata (PDAs), a way for implementing the strategy as a parsing algorithm. While previous studies on the arc-eager left-corner PDAs pay less attention on its theoretical properties beyond its asymptotic behavior, in Section 2.2.4, we present a detailed, thorough analysis on the properties of the presented PDA as it plays an essential role in our exploration in the following chapters. Although we carefully design the left-corner PDA as the realization of the presented strategy, as we see later, this algorithm differs from the one previously explained as the left-corner PDAs in the literature (Resnik, 1992; Johnson, 1998a). This difference is important for us. In Section 2.2.5 we discuss why this discrepancy occurs, as well as why we do not take the standard formalization. Finally in Section 2.2.6 we summarize the psycholinguistic relevance of the presented algorithms.

2.2.1 Center-embedding

We first define some additional notations related to CFGs that we introduced in Section 2.1.1. Let us assume a CFG $G = (N, \Sigma, P, S)$. Then each symbol used below has the following meaning:

- A, B, C, \dots are nonterminal symbols;
- v, w, x, \dots are strings of terminal symbols, e.g., $v \in \Sigma^*$;

trees, we ignore those labels in Figure 2.9.

³Johnson-Laird (1983) introduced his left-corner parser as a cognitively plausible human parser but it has been pointed out that his parser is actually not arc-eager but arc-standard (Resnik, 1992), which is (at least) not relevant to a human parser.

- $\alpha, \beta, \gamma, \dots$ are strings of terminal or nonterminal symbols, e.g., $\alpha \in (N \cup \Sigma)^*$.

In the following, we define the notion of center-embedding using left-most *derives* relation \Rightarrow_{lm} though it is also possible to define with right-most one. \Rightarrow_{lm}^* denotes derivation in zero or more steps while \Rightarrow_{lm}^+ denotes derivation in one or more steps. $\alpha \Rightarrow_{lm}^* \beta$ means β can be derived from α by applying a list of rules in left-most order (always expanding the current left-most nonterminal). In this order, the derivation with nonterminal symbols followed by terminal symbols, i.e., $S \Rightarrow_{lm}^+ \alpha Av$ does not appear.

For simplicity, we assume the CFG is in CNF. It is possible to define center-embedding for general CFGs but notations are more involved, and it is sufficient for discussing our extension for dependency grammars.

Center-embedding can be characterized by the specific branching pattern found in a CFG parse, which we define precisely below. We note that the notion of center-embedding could be defined in a different way. In fact, as we describe later, the existence of several variants of arc-eager left-corner parsers is relevant to this arbitrariness for the definition of center-embedding. We postpone the discussion of this issue until Section 2.2.5.

Definition 2.1. A CFG parse involves center-embedding if the following derivation is found in it:

$$S \Rightarrow_{lm}^* v \underline{A} \alpha \Rightarrow_{lm}^+ vw \underline{B} \alpha \Rightarrow_{lm} vw \underline{C} D \alpha \Rightarrow_{lm}^+ vw x D \alpha; \quad |x| \geq 2,$$

where the underlined symbol indicates that that symbol is expanded by the following \Rightarrow . The condition $|x| \geq 2$ means the constituent rooted at C must comprise of more than one word.

Figure 2.10(a) shows an example of the branching pattern. The pattern always begins with right branching edges, which are indicated by $v \underline{A} \beta \Rightarrow_{lm}^+ vw \underline{B} \beta$. Then the center-embedding is detected if some B is found which has a left child that constitutes a span larger than one word (e.g., C). The final condition of the span length means the embedded subtree (rooted at C) has a right child. This *right* \rightarrow *left* \rightarrow *right* pattern is the distinguished branching pattern in center-embedding.

By detecting a series of these zig-zag patterns recursively, we can measure the *degree* of center-embedding in a given parse. Formally,

Definition 2.2. If the following derivation is found in a CFG parse:

$$\begin{aligned} S &\Rightarrow_{lm}^* v \underline{A} \alpha \Rightarrow_{lm}^+ vw_1 \underline{B}_1 \alpha \Rightarrow_{lm}^+ vw_1 \underline{C}_1 \beta_1 \alpha \\ &\Rightarrow_{lm}^+ vw_1 w_2 \underline{B}_2 \beta_1 \alpha \Rightarrow_{lm}^+ vw_1 w_2 \underline{C}_2 \beta_2 \beta_1 \alpha \\ &\Rightarrow_{lm}^+ \dots \\ &\Rightarrow_{lm}^+ vw_1 \dots w_{m'} \underline{B}_{m'} \beta_{m'-1} \dots \beta_1 \alpha \Rightarrow_{lm}^+ vw_1 \dots w_{m'} \underline{C}_{m'} \beta_{m'} \beta_{m'-1} \dots \beta_1 \alpha \\ &\Rightarrow_{lm}^+ vw_1 \dots w_{m'} x \beta_{m'} \beta_{m'-1} \dots \beta_1 \alpha; \quad |x| \geq 2, \end{aligned} \tag{2.1}$$

the degree of center-embedding in it is the maximum value m among all possible values of m' (i.e., $m \geq m'$).

Each line in Eq. 2.1 corresponds to the detection of additional embedding, except the last line that checks the length of the most embedded constituent. Figures 2.10(b) and 2.10(c) show examples

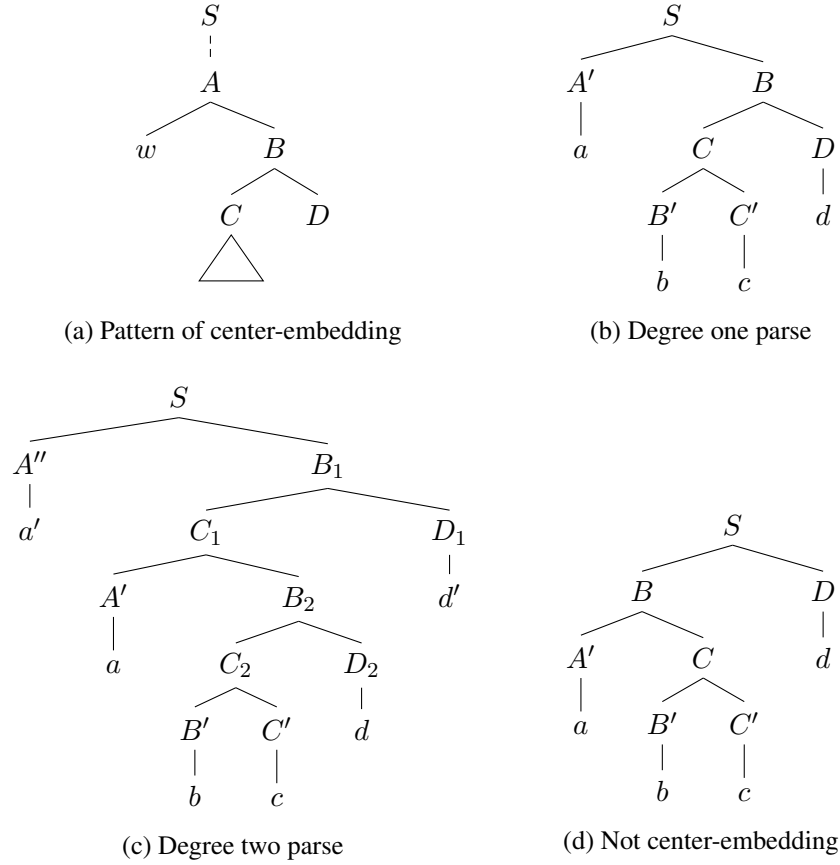


Figure 2.10: A parse involves center-embedding if the pattern in (a) is found in it. (b) and (c) are the minimal patterns with degree one and two respectively. (d) is the symmetry of (b) but we regard this as not center-embedding.

of degree one and two parses, respectively. These are the *minimal* parses for each degree, meaning that degree two occurs only when the sentence length ≥ 6 . Note that the form of the last transform in the first line (i.e., $vw_1 \underline{B_1} \alpha \Rightarrow_{lm}^+ vw_1 \underline{C_1} \beta_1 \alpha$) does not match to the one in Definition 2.1 (i.e., $vw \underline{B} \alpha \Rightarrow_{lm} vw \underline{C} D \alpha$). This modification is necessary because the first left descendant of B_1 in Eq. 2.1 is not always the starting point of further center-embedding.

Other notes about center-embedding are summarized as follows:

- It is possible to calculate the depth m by just traversing every node in a parse once in a left-to-right, depth-first manner. The important observation for this algorithm is that the value m' in Eq. 2.1 is deterministic for each node, suggesting that we can fill the depth of each node top-down. We then pick up the maximum depth m among the values found at terminal nodes.
- In the definitions above, the pattern of center-embedding always starts with a right directed edge. This means the similar, but opposite pattern found in Figure 2.10(d) is not center-embedding. Left-corner parsing that we will introduce next also distinguish only the pattern

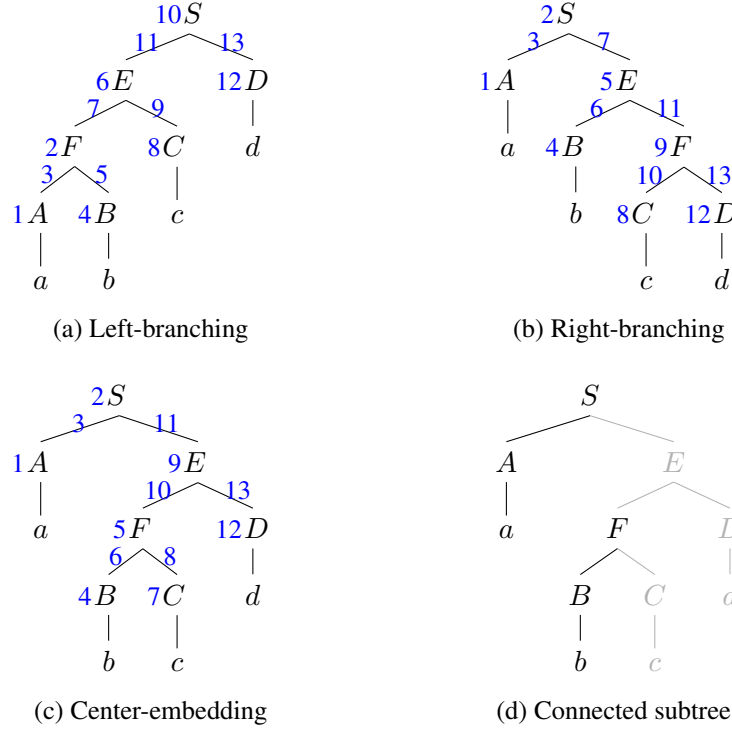


Figure 2.11: (a)–(c) Three kinds of branching structures with numbers on symbols and arcs showing the order of recognition with a left-corner strategy. (d) a partial parse of (c) using a left-corner strategy just after reading symbol b , with gray edges and symbols showing elements not yet recognized; The number of connected subtrees here is 2.

in Figure 2.10(b), not Figure 2.10(d).

2.2.2 Left-corner parsing strategy

A parsing strategy is a useful abstract notion for characterizing the properties of a parser and gaining intuition into parser behavior. Formally, it can be understood as a particular mapping from a CFG to the push-down automata that generate the same language (Nederhof and Satta, 2004a). Here we follow Abney and Johnson (1991) and consider a parsing strategy as a specification of the order that each node and arc on a parse is recognized during parsing. The corresponding push-down automata can then be understood as the device that provides the operational specification to realize such specific order of recognition, as we describe in Section 2.2.3.

We first characterize left-corner parsing with a parsing strategy to discuss its notable behavior for center-embedded structures. The left-corner parsing strategy is defined by the following order of recognizing nodes and arcs on a parse tree:

1. A node is recognized when the subtree of its first (left most) child has been recognized.
2. An arc is recognized when two nodes it connects have been recognized.

We discuss the property of the left-corner strategy based on its behavior on three kinds of distinguished tree structures called left-branching, right-branching, and center-embedding, each shown in Figure 2.11. The notable property of the left-corner strategy is that it generates disconnected tree fragments only on a center-embedded structure as shown in Figure 2.11. Specifically, in Figure 2.11(c), which is center-embedding, after reading b it reaches 6 but a and b cannot be connected at this point. It does not generate such fragments for other structures; e.g., for the right-branching structure (Figure 2.11(b)), it reaches 7 after reading b so a and b are connected by a subtree at this point. The number of tree fragments grows as the degree of center-embedding increases.

As we describe later, the property of the left-corner strategy is appealing from a psycholinguistic viewpoint. Before discussing this relevance, which we summarize in Section 2.2.6, in the following we will see how this strategy can be actually realized as the parsing algorithm first.

2.2.3 Push-down automata

We now discuss how the left-corner parsing strategy described above is implemented as a parsing algorithm, in particular as push-down automata (PDAs), the common device to define a parsing algorithm following a specific strategy. As we mentioned, this algorithm is not exactly the same as the one previously proposed as the left-corner PDA (Resnik, 1992; Johnson, 1998a), which we summarize in Section 2.2.5.

PDAs assume a CFG, and specify how to build parses with that grammar given an input sentence. Note that for simplicity we only present algorithms specific for CFGs in CNF, although both presented algorithms can be extended for general CFGs.

Notations We define a PDA as a tuple $(\Sigma, Q, q_{init}, q_{final}, \Delta)$ where Σ is an alphabet of input symbols (words) in a CFG, Q is a finite set of stack symbols (items), including the initial stack symbol q_{init} and the final stack symbol q_{final} , and Δ is a finite set of transitions. A transition has the form $\sigma_1 \xrightarrow{a} \sigma_2$ where $\sigma_1, \sigma_2 \in Q^*$ and $a \in \Sigma \cup \{\varepsilon\}$; ε is an empty string. This can be applied if the stack symbols σ_1 are found to be the top few symbols of the stack and a is the first symbol of the unread part of the input. After such a transition, σ_1 is replaced with σ_2 and the next input symbol a is treated as having been read. If $a = \varepsilon$, the input does not proceed. Note that our PDA does not explicitly have a set of states; instead, we encode each state into stack symbols for simplicity as Nederhof and Satta (2004b).

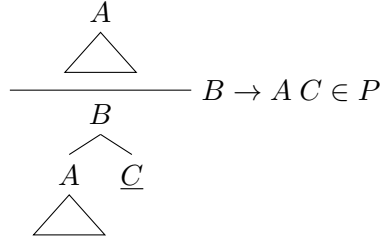
Given a PDA and an input sentence of length n , a *configuration* of a PDA is a pair (σ, i) where a stack $\sigma \in Q^*$ and i is an input position $1 \leq i \leq n$, indicating how many symbols are read from the input. The initial configuration is $(q_{init}, 0)$ and the PDA *recognizes* a sentence if it reaches (q_{final}, n) after a finite number of transitions.

PDA We develop a left-corner PDA to achieve the recognition order of nodes and arcs by the left-corner strategy that we formulated in Section 2.2.2. In our left-corner PDA, each stack symbol is either a nonterminal $A \in N$, or a pair of nonterminals A/B , where $A, B \in N$. A/B is used for representing an *incomplete* constituent, waiting for a subtree rooted at B being substituted. In this algorithm, q_{init} is an empty stack symbol ε while q_{final} is the stack symbol S , the start symbol of a given CFG.

| Name | Transition | Condition |
|-------------|---------------------------------------|---------------------------|
| SHIFT | $\varepsilon \xrightarrow{a} A$ | $A \rightarrow a \in P$ |
| SCAN | $A/B \xrightarrow{a} A$ | $B \rightarrow a \in P$ |
| PREDICTION | $A \xrightarrow{\varepsilon} B/C$ | $B \rightarrow A C \in P$ |
| COMPOSITION | $A/B C \xrightarrow{\varepsilon} A/D$ | $B \rightarrow C D \in P$ |

Figure 2.12: The set of transitions in a push-down automaton that parses a CFG (N, Σ, P, S) with the left-corner strategy. $a \in \Sigma$; $A, B, C, D \in N$. The initial stack symbol q_{init} is the start symbol of the CFG S , while the final stack symbol q_{final} is an empty stack symbol ε .

PREDICTION:



COMPOSITION:

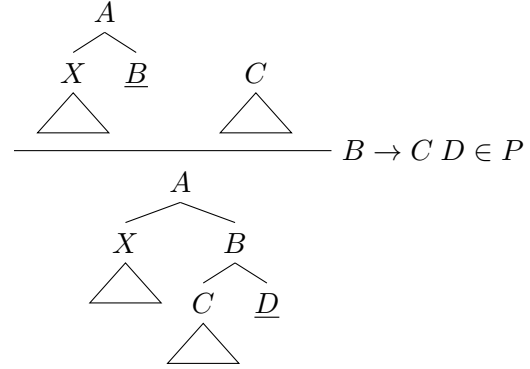


Figure 2.13: Graphical representations of inference rules of PREDICTION and COMPOSITION defined in Figure 2.12. An underlined symbol indicates that the symbol is predicted top-down.

Figure 2.12 lists the set of transitions in this PDA. PREDICTION and COMPOSITION are the key operations for achieving the left-corner strategy. Specifically, PREDICTION operation first recognizes a parent node of a subtree (rooted at A) bottom-up, and then predicts its sibling node top-down. This is graphically explained in Figure 2.13. We notice that this operation just corresponds to the policy 1 of the strategy about the order of recognizing nodes.⁴ The policy 2 about the order of connecting nodes is also essential, and it is realized by another key operation of COMPOSITION. This operation involves two steps. First, it performs the same prediction operation as PREDICTION for the top stack symbol. It is C in Figure 2.12, and the result is B/D , i.e., a subtree rooted at B predicting the sibling node D . It then connects this subtree and the second top subtree, i.e., A/B . This is done by matching two identical nodes of different views, i.e., top-down predicted node B in A/B and bottom-up recognized node B in B/D . This matching operation is the key for achieving the policy 2, which demands that two recognized nodes be connected immediately. In COMPOSITION, these two nodes are A , which is already recognized, and B , which is just recognized bottom-up by

⁴ Though the strategy postpones the recognition of the sibling node, we can interpret that the predicted sibling (i.e., C) by PREDICTION is still not recognized. It is recognized by SCAN or COMPOSITION, which introduce the same node bottom-up and matches two nodes, i.e., the top-down predicted node and the bottom-up recognized node.

| Step | Action | Stack | Read symbol |
|------|-------------|---------------|-------------|
| | | ε | |
| 1 | SHIFT | A' | a |
| 2 | PREDICT | S/B | |
| 3 | SHIFT | $S/B B'$ | b |
| 4 | PREDICT | $S/B C/C'$ | |
| 5 | SCAN | $S/B C$ | c |
| 6 | COMPOSITION | S/D | |
| 7 | SCAN | S | d |

Figure 2.14: An example of parsing process by the left-corner PDA to recover the parse in Figure 2.10(b) given an input sentence $a b c d$. It is step 4 that occurs stack depth two after a reduce transition.

$$\begin{array}{ll}
 S \rightarrow A' B & A' \rightarrow a \\
 B \rightarrow C D & B' \rightarrow b \\
 C \rightarrow B' C' & C' \rightarrow c \\
 & D \rightarrow d
 \end{array}$$

Figure 2.15: A CFG that is parsed with the process in Figure 2.14.

the first prediction step.⁵

In the following, we distinguish two kinds of transitions in Figure 2.12: SHIFT and SCAN operations belong to *shift* transitions⁶, as they proceed the input position of the configuration. This is not the case in PREDICTION and COMPOSITION, and we call them *reduce* transitions.

The left-corner strategy of Abney and Johnson (1991) has the property that the maximum number of unconnected subtrees during enumeration equals the degree of center-embedding. The presented left-corner PDA is an implementation of this strategy and essentially has the same property; that is, its maximum stack depth during parsing is equal the degree of center-embedding of the resulting parse. The example of this is shown next, while the formal discussion is provided in Section 2.2.4.

Example Figure 2.14 shows an example of parsing process given a CFG in Figure 2.15 and an input sentence $a b c d$. The parse tree contains one degree of center-embedding found in Figure 2.10(b), and this is illuminated in Figure 2.14 with the appearances of stack depth of two, in particular before reading symbol c , which exactly corresponds to the step 4 on the Figure 2.11(c).

2.2.4 Properties of the left-corner PDA

In this section, we formally establish the connection between the left-corner PDA and the center-embeddedness of a parse. The result is also essential when discussing the property of our extended

⁵ As mentioned in footnote 4, we regard the predicted node B in A/B as not yet being recognized.

⁶ We use small caps to refer to a specific action, e.g., SHIFT, while “shift” refers to an action type.

algorithm for dependency grammars presented in Chapter 4; see Section 4.3.4 for details.

The following lemmas describe the basic properties of the left-corner PDA, which will be the basis in the further analysis.

Lemma 2.1. *In a sequence of transitions to arrive the final configuration (q_{final}, n) of the left-corner PDA (Figure 2.12), shift (i.e., SHIFT or SCAN) and reduce (i.e., PREDICTION or COMPOSITION) transitions occur alternately.*

Proof. Reduce transitions are only performed when the top symbol of the stack is *complete*, i.e., of the form A . Then, since each reduce transition makes the top symbol of the stack incomplete, two consecutive reduce transitions are not applicable. Conversely, shift transitions make the top stack symbol complete. We cannot perform SCAN after some shift transition, since it requires an incomplete top stack symbol. If we perform SHIFT after a shift transition, the top two stack symbols become complete, but we cannot combine these two symbols since the only way to combine two symbols on the stack is COMPOSITION, while it requires the second top symbol to be incomplete. ■

Lemma 2.2. *In the left-corner PDA, after each reduce transition, every item remained on the stack is an incomplete stack symbol of the form A/B .*

Proof. From Lemma 2.1, a shift action is always followed by a reduce action, and vice versa. We call a pair of some shift and reduce operations a *push* operation. In each push operation, a shift operation adds at most one stack symbol on the stack, which is always replaced with an incomplete symbol by the followed reduce transition. Thus after a reduce transition no complete symbol remains on the stack. ■

We can see that transitions in Figure 2.14 satisfy these conditions. Intuitively, the existence of center-embedding is indicated by the accumulated incomplete symbols on the stack, each of which corresponds to each line on the derivation in Eq. 2.1. This is formally stated as the following theorem, which establishes the connection between the stack depth of the left-corner PDA and the degree of center-embedding.

Theorem 2.1. *Given a CFG parse, its degree of center-embedding is equal to the maximum value of the stack depth after a reduce transition minus one for recognizing that parse on the left-corner PDA.*

For example, for a CFG parse with one degree of center-embedding, the maximum stack depth after a reduce transition is two, which is indicated at step 4 in Figure 2.14. We leave the proof of this theorem in Appendix A.

Note that Theorem 2.1 says nothing about the stack depth after a shift transition, which generally is not equal to the degree of center-embedding. We discuss this issue more when presenting the algorithm for dependency grammars; see Section 4.3.4.

| Name | Transition | Condition |
|-------------|---------------------------------------|---------------------------|
| SHIFT | $A \xrightarrow{a} A-B$ | $B \rightarrow a \in P$ |
| SCAN | $A \xrightarrow{a} \varepsilon$ | $A \rightarrow a \in P$ |
| PREDICTION | $A-B \xrightarrow{\varepsilon} A-C D$ | $C \rightarrow B D \in P$ |
| COMPOSITION | $A-B \xrightarrow{\varepsilon} C$ | $A \rightarrow B C \in P$ |

Figure 2.16: A set of transitions in another variant of the left-corner PDA appeared in Resnik (1992). $a \in \Sigma$; $A, B, C, D \in N$. Differently from the PDA in Figure 2.12, the initial stack symbol q_{init} is S while q_{final} is an empty stack symbol ε .



Figure 2.17: Stack symbols of the left-corner PDA of Figure 2.16. Both trees correspond to symbol $A-B$ where A is the current goal while B is the recognized nonterminal. Note that A may be a right descendant of another nonterminal (e.g., X), which dominates a larger subtree.

2.2.5 Another variant

We now present another variant of the left-corner PDA appeared in the literature (Resnik, 1992; Johnson, 1998a). We will see that this algorithm has a different property with respect to the stack depth and the degree of center-embedding than Theorem 2.1. In particular, this difference is relevant to the structures that are recognized as center-embedding for the algorithm, which has not been precisely discussed so far; Schuler et al. (2010) give comparison of two algorithms but from a different perspective.

Figure 2.16 shows the list of possible transitions in this variant. The crucial difference between two PDAs is in the form of initial and final stack symbols. That is, in this PDA the initial stack symbol $q_{initial}$ is S , while q_{final} is an empty symbol ε , which are opposite in the PDA that we discussed so far (Section 2.2.3).

Also in this variant, the form of stack symbols is different. Instead of A/B , which represents a subtree waiting for B , it has $A-B$, which means that B is the *left-corner* in a subtree rooted at A , and has been already recognized. In other words, A is the current goal, which the PDA tries to build, while B represents a finished subtree. This is schematically shown in Figure 2.17(a).

Parsing starts with $q_{initial} = S$, which immediately changes to $S-A$ where $A \rightarrow a$, and $a \in \Sigma$ is the initial token of the sentence. PREDICTION is similar to the one in our variant: It expands the currently recognized structure, and also predicts the sibling symbol (i.e., D), which becomes a new goal symbol. COMPOSITION looks very different, but has the similar sense of transition.

| Step | Action | Stack | Read symbol |
|------|-------------|------------|-------------|
| | | S | |
| 1 | SHIFT | $S-A'$ | a |
| 2 | COMPOSITION | B | |
| 3 | SHIFT | $B-B'$ | b |
| 4 | PREDICT | $B-C \ C'$ | |
| 5 | SCAN | $B-C$ | c |
| 6 | COMPOSITION | D | |
| 7 | SCAN | ϵ | d |

Figure 2.18: Parsing process of the PDA in Figure 2.16 to recover the parse in Figure 2.10(b) given the CFG in Figure 2.15 and an input sentence $a b c d$. The stack depth keeps one in every step after a shift transition.

In the symbol A/B , A is not limited to S , in which case A is some right descendant of another nonterminal, as depicted in Figure 2.17(b). The sense of COMPOSITION in Figure 2.16 is that we finish recognition of the left subtree of A (i.e., the tree rooted at B) and change the goal symbol to C , the sibling of B . If we consider this transition in the form of Figure 2.17(b), it looks similar to the one in Figure 2.12; that is, the corresponding transition in our variant is $X/A \ B \xrightarrow{\epsilon} C$. Instead, in the current variant, the root nonterminal of a subtree X is not kept on the stack, and the goal symbol is moved from top to bottom. This is the reason why the final stack symbol q_{final} is empty. The final goal for the PDA is always the preterminal for the last token of the sentence, which is then finally removed by SCAN.

Example This PDA has slightly different characteristics in terms of stack depth and the degree of center-embedding, which we point out here with some examples. In particular, it regards the parse in Figure 2.10(d) as singly (degree one) center-embedded, while the one in Figure 2.10(b) as not center-embedded. That is, it has just the opposite properties to the PDA that we discussed in Section 2.2.3.

We first see how the situation changes for the CFG that we gave an example in Figure 2.14, which analyzed the parse in Figure 2.10(b). See Figure 2.18. Contrary to our variant, this PDA has the property that its stack depth after some *shift* transitions increases as the degree of center-embedding increases.⁷ In this case, these are steps 3, 5, and 7, all of which has a stack with only one element. The main reason why it does not increase the stack depth is in the first COMPOSITION operation, which changes the stack symbol to B . After that, since the outside structure of B is already processed, the remaining tree looks just like left-branching, which the left-corner PDA including this variant processes without increasing the stack depth.

On the other hand, for the parse in Figure 2.10(d), this PDA increases the stack depth as simulated in Figure 2.19. At step 2, the PDA introduces new goal symbol C , which remains on the stack

⁷ This again contrasts with our variant (Theorem 2.1). This is because in the PDA in 2.16, new stack element is introduced with a reduce transition (i.e., PREDICTION), and center-embedding is detected with the followed SHIFT, which does not decrease the stack depth. In our variant, on the other hand, new stack element is introduced by SHIFT. Center-embedding is detected if this new element remains on the stack after a reduce transition (by PREDICTION).

| Step | Action | Stack | Read symbol |
|------|-------------|---------------|-------------|
| | | S | |
| 1 | SHIFT | $S-A'$ | a |
| 2 | PREDICTION | $S-B C$ | |
| 3 | SHIFT | $S-B C-B'$ | b |
| 4 | COMPOSITION | $S-B C'$ | |
| 5 | SCAN | $S-B$ | c |
| 6 | COMPOSITION | D | |
| 7 | SCAN | ε | d |

Figure 2.19: Parsing process of the PDA in Figure 2.16 to recover the parse in Figure 2.10(d). The stack depth after a shift transition increases at step 3.

after the followed SHIFT. This is the pattern of transitions with which this PDA increase its stack depth, and it occurs when processing the zig-zag patterns starting from left edges, not right edges as in our variant.

Discussion We have pointed out that there are two variants of (arc-eager) left-corner PDAs, which suffer from slightly different conditions under which their stack depth increases. From an empirical point of view, the only common property is its asymptotic behavior. That is, both linearly increase the stack depth as the degree of center-embedding increases. The difference is rather subtle, i.e., the condition of beginning center-embedding (left edges or right edges).

Historically, the variant introduced in this section (Figure 2.16) has been thought as the realization of the left-corner PDA (Resnik, 1992; Johnson, 1998a). However, as we have seen, if we base development of the algorithm on the parsing strategy (Section 2.2.2), our variant can be seen as the correct implementation of it, as only our variant preserves the transparent relationship between the stack depth and the disconnected trees generated during enumeration by the strategy.

Resnik (1992) did not design the algorithm based on the parsing strategy, but from an existing arc-standard left-corner PDA (Rosenkrantz and Lewis, 1970; Johnson-Laird, 1983), which also accepts an empty stack symbol as the final configuration. His main argument is that the arc-eager left-corner PDA can be obtained by introducing a COMPOSITION operation, which does not exist in the arc-standard PDA. Interestingly, there is another variant of the arc-standard PDA (Nederhof, 1993), which instead accepts the S symbol⁸. If we extend this algorithm by introducing COMPOSITION, we get very similar algorithm to the one we presented in Section 2.2.3 with the same stack depth property.

Thus, we can conclude that Resnik's argument is correct in that a left-corner PDA can be *arc-eager* by adding composition operations, but depending on which arc-standard PDA we employ as the basis, the resulting arc-eager PDA may have different characteristics in terms of stack depth. In particular, the initial and final stack configurations are important. If the based arc-standard PDA accepts the empty stack symbol as in Rosenkrantz and Lewis (1970), the corresponding arc-eager

⁸To be precise, the stack item of Nederhof (1993) is a dotted rule like $[S \rightarrow NP \bullet VP]$ and parsing finishes with an item of the form $[S \rightarrow \alpha \bullet]$ with some α .

PDA regards the pattern beginning with right edges as center-embedding. The direction becomes opposite if we start from the PDA that accepts the non-empty stack symbol as in Nederhof (1993).

Our discussion in the following chapters is based on the variant we presented in Section 2.2.3, which is relevant to Nederhof (1993). However, we do not make any claims such that this algorithm is superior to the variant we introduced in this section. Both are correct arc-eager left-corner PDAs, and we argue that the choice is rather arbitrary. This arbitrariness is further discussed next, along with the limitation of both approaches as the psycholinguistic models.

Finally, our variant of the PDA in Section 2.2.3 has been previously presented in Schuler et al. (2010) and van Schijndel and Schuler (2013), though they do not mention the relevance of the algorithm to the parsing strategy. Their main concern is in the psychological plausibility of the parsing model, and they argue that this variant is more plausible due to its inherent bottom-up nature (not starting from the predicted S symbol). They do not point out the difference of two algorithms in terms of the recognized center-embedded structures as we discussed here.

2.2.6 Psycholinguistic motivation and limitation

We finally summarize left-corner parsing and relevant theories in the psycholinguistics literature. One well known observation about human language processing is that the sentences with multiply center-embedded constructions are quite difficult to understand, while left- and right-branching constructions seem to cause no particular difficulty (Miller and Chomsky, 1963; Gibson, 1998).

- (2) a. # The reporter [who the senator [who Mary met] attacked] ignored the president.
- b. Mary met the senator [who attacked the reporter [who ignored the president]].

The sentence (2a) is an example of a center-embedded sentence while (2b) is a right-branching sentence. This observation matches the behavior of left-corner parsers, which increase its stack depth in processing center-embedded sentences only, as we discussed above.

It has been well established that center-embedded structures are a generally difficult construction (Gibson, 1998; Chen et al., 2005), and this connection between left-corner parsers and human behaviors motivated researchers to investigate left-corner parsers as an approximation of human parsers (Roark, 2001; Schuler et al., 2010; van Schijndel and Schuler, 2013). The most relevant theory in psycholinguistics that accounts for the difficulty of center-embedding is the one based on the *storage cost* (Chen et al., 2005; Nakatani and Gibson, 2010; Nakatani and Gibson, 2008), i.e., the cost associated with keeping incomplete materials in memory.⁹ For example, Chen et al. (2005) and Nakatani and Gibson (2010) find that people read more center-embedded sentences more slower than less center-embedded sentences, in particular when entering new embedded clauses, through their reading time experiments of English and Japanese, respectively. This observation suggests that there exists some sort of *storage* component in human parsers, which is consumed when processing more nested structures, as in the stack of left-corner parsers.

⁹ Another explanation for this difficulty is retrieval-based accounts such as the *integration cost* (Gibson, 1998; Gibson, 2000) in the dependency locality theory. We do not discuss this theory since the connection between the integration cost and the stack depth of left-corner parsers is less obvious, and it has been shown that the integration cost itself is not sufficient to account for the difficulty of center-embedding (Chen et al., 2005; Nakatani and Gibson, 2010).

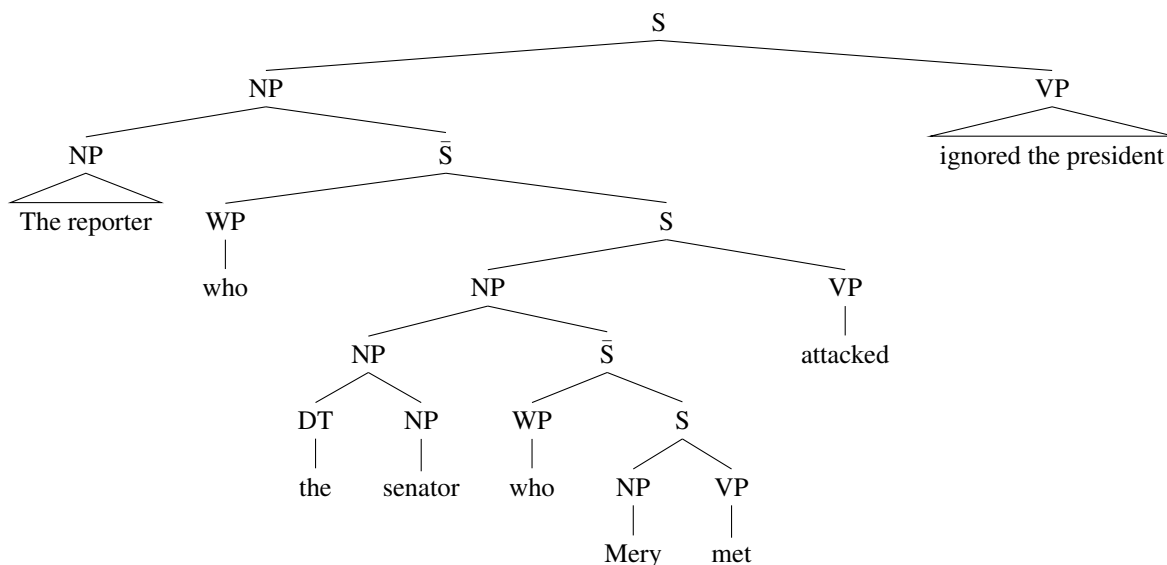


Figure 2.20: The parse of the sentence (2a).

However, as we claimed in Section 1.2, our main goal in this thesis is not to deepen understanding of the mechanism of human sentence processing. One reason of this is that there are some discrepancies between the results in the articles cited above and the behavior of our left-corner parser, which we summarize below. Another, and perhaps more important limitation of left-corner parsers as an approximation of human parsers is that it cannot account for the sentence difficulties not relevant to center-embedding, such as the garden path phenomena:

(3) # The horse raced past the barn fell,

in which people feel difficulty at the last verb *fell*. Also there exist some cases in which nested structures *do* facilitate comprehension, known as *anti-locality* effects (Konieczny, 2000; Shravan Vasishth, 2006). These can be accounted for by another, non-memory-based theory called expectation-based account (Hale, 2001; Levy, 2008), which is orthogonal in many aspects to the memory-based account (Jaeger and Tily, 2011). We do not delve into those problems further and in the following we focus on the issues of the former mentioned above, which is relevant to our definition of center-embedding as well as the choice of the variant of left-corner PDAs (Section 2.2.5).

Discrepancies in definitions of center-embedding We argue here that sometimes the stack depth of our left-corner parser *underestimates* the storage cost for some center-embedded sentences in which linguists predict greater difficulty for comprehension. More specifically, though Chen et al. (2005) claims the sentence (2a) is *doubly* center-embedded, our left-corner parser recognizes this is *singly* center-embedded, as its parse does not contain the zig-zag pattern in Figure 2.10(c) (but in Figure 2.10(b)). Figure 2.20 shows the parse. This discrepancy occurs due to our choice for the

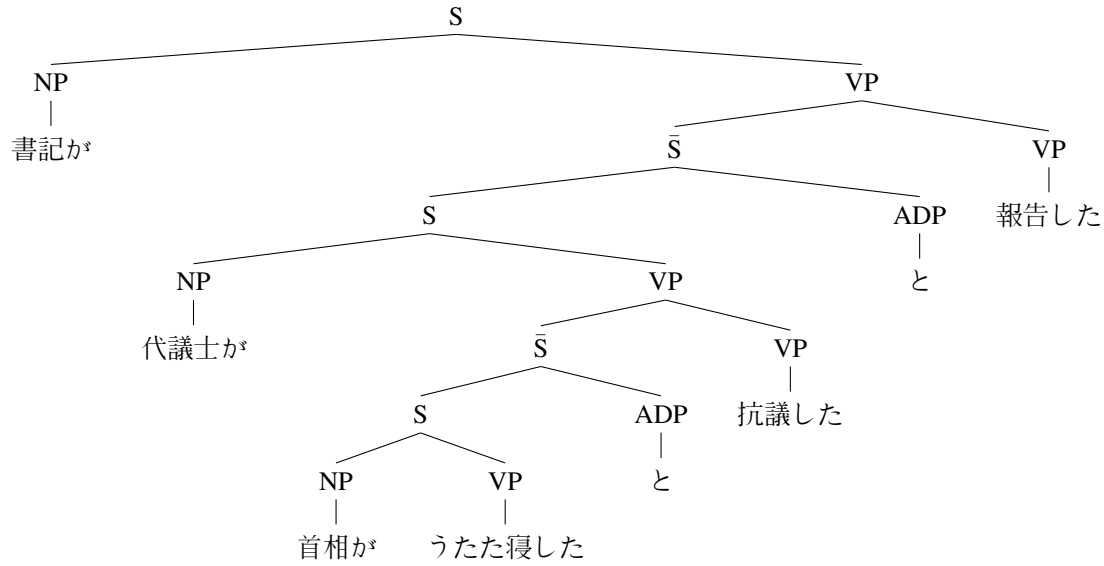


Figure 2.21: The parse of the sentence (5).

definition of center-embedding discussed in Section 2.2.1. In our definition (Definition 2.1), center-embedding always starts with a right edge. In the case like Figure 2.20, two main constituents “The reporter ... attacked” and “ignored the president” are connected with a left edge, and this is the reason why our definition of center-embedding as well as our left-corner parser predicts that this parse is singly nested.

Here we note that although our left-corner parser underestimates the center-embeddedness in some cases, it correctly estimates the relative difficulty of sentence (2a) compared to less nested sentences below.

- (4) a. The senator [who Mary met] ignored the president.
 b. The reporter ignored the president.

The problem is that both sentences above are recognized as not center-embedded although some literature in psycholinguistics (e.g., Chen et al. (2005)) assumes it is singly center-embedded.

書記が

However, this mismatch does not mean that our left-corner parser always underestimates the predicted center-embeddedness by linguists. We give further examples below to make explicit the points.

- As the example below (Nakatani and Gibson, 2008) indicates, often in the parse of a Japanese sentence the degree of center-embedding matches the prediction by linguists.
- (5) # 書記が [代議士が [首相が うたた寝した と] 抗議した と] 報告した
 secretary-nom [congressman-nom [prime minister-nom dozed comp] protested comp]
 reported

The secretary reported that the congressman protested that the prime minister had dozed.

The parse is shown in Figure 2.21, which contains the pattern in Figure 2.10(c). This is because two constituents “書記が” and “代議士が ... 報告した” are connected with a right edge in this case.

- This observation may suggest that our left-corner parser always underestimates the degree of center-embedding for specific languages, e.g., English. However, this is not generally true since we can make an English example in which two predictions are consistent, as in Japanese sentence, e.g., by making the sentence (2) as a large complement as follows:

(6) # He said [the reporter [who the senator [who Mary met] attacked] ignored the president].

In the example, “He said” does not cause additional embedding, as the constituent “the reporter ... president” is not embedded internally, and thus linguists predict that this is still doubly center-embedded. On the other hand, the parse now involves the pattern in Figure 2.10(c), suggesting that the predictions are consistent in this case.

The point is that since our left-corner parser (PDA) only regards the pattern starting from right edges as center-embedding, it underestimates the prediction by linguists when the direction of outermost edge in the parse is left, as in Figure 2.20. Though there might be some language specific tendency (e.g., English sentences might be often underestimated) we do not make such claims here, since the degree of center-embedding in our definition is determined purely in terms of the tree structure, as indicated by sentence (6). We perform the relevant empirical analysis on treebanks in Chapter 4.

From the psycholinguistics viewpoint, this discrepancy might make our empirical studies in the following chapters less attractive. However, as we noted in Section 1.2, our central motivation is rather to capture the universal constraint that every language may suffer from, though is computationally tractable, which we argue does not necessarily reflect correctly the difficulties reported by psycholinguistic experiments.

As might be predicted, the results so far become opposite if we employ another variant of PDA that we formulated in Section 2.2.5, in which the stack depth increases on the pattern starting from left edges, as in Figure 2.10(d). This variant of PDA estimates that the degree of center-embedding on the parse in Figure 2.20 will be two, while that of Figure 2.21 will be one. This highlights that the reason of the observed discrepancies is mainly due to the computational tractability: We can develop a left-corner parser so that its stack depth increases on center-embedded structures indicated by some zig-zag patterns, which are always starting from left (the variant of Resnik (1992)), or right (our variant). However, from an algorithm perspective, it is hard to allow both left and right directions, and this is the assumption of psycholinguists.

Again, we do argue that our choice for the variant of the left-corner PDA is rather arbitrary. This choice may impact the empirical results in the following chapters, where we examine the relationships between parses on the treebanks and the incurred stack depth. In the current study, we do not empirically compare the behaviors of two PDAs, which we leave as one of future investigations.

2.3 Learning Dependency Grammars

In this section we will summarize several basic ideas about learning and parsing of dependency grammars. The dependency model with valence (Klein and Manning, 2004) is the most popular model for unsupervised dependency parsing, which will be the basis of our experiments in Chapter 5. We formalize this model in Section 2.3.6 as a special instance of split bilexical grammars (Section 2.3.5). Before that, this section first reviews some preliminaries on a learning mechanism, namely, probabilistic context-free grammars (Section 2.3.1), chart parsing (Section 2.3.2), and parameter estimation with the EM algorithm (Section 2.3.3).

2.3.1 Probabilistic context-free grammars

Here we start the discussion with probabilistic context-free grammars (PCFGs) because they will allow use of generic parsing and parameter estimation methods that we describe later. However, we note that for the grammar to be applied these algorithms, the grammar should not necessarily be a PCFG. We will see that in fact split bilexical grammars introduced later cannot always be formulated as a correct PCFG. Nevertheless, we begin this section with the discussion of PCFGs mainly because:

- the ideas behind chart parsing algorithms (Sections 2.3.2 and 2.3.3) can be best understood with a simple PCFG; and
- we can obtain a natural generalization of these algorithms to handle a special class of grammars (not PCFGs) including split bilexical grammars. We will describe the precise condition for a grammar to be applied these algorithms later.

Formally a PCFG is a tuple $G = (N, \Sigma, P, S, \theta)$ where (N, Σ, P, S) is a CFG (Section 2.1.1) and θ is a vector of non-negative real values indexed by production rules P such that

$$\sum_{A \rightarrow \beta \in P_A} \theta_{A \rightarrow \beta} = 1, \quad (2.2)$$

where $P_A \subset P$ is a collection of rules of the form $A \rightarrow \beta$. We can interpret $\theta_{A \rightarrow \beta}$ as the conditional probability of choosing a rule $A \rightarrow \beta$ given that the nonterminal being expanded is A .

With this model, we can calculate the score (probability) of a parse as the product of rules appeared on that. Let a parse be z that contains rules r_1, r_2, \dots ; then the probability of z under the given PCFG is

$$P(z|\theta) = \prod_{r_i \in z} \theta_{r_i} \quad (2.3)$$

$$= \prod_{r \in P} \theta_r^{f(r,z)}, \quad (2.4)$$

where $f(r, z)$ is the number of occurrences of a rule r in z .

We can also interpret a PCFG as a directed graphical model that defines a distribution over CFG parses. The generative process is described as follows: Starting at S (start symbol), it chooses to

apply a rule $S \rightarrow \beta$ with probability $\theta_{S \rightarrow \beta}$; β defines the symbols of the children, which are then expanded recursively to generate their subtrees. This process stops when all the leaves of the tree are terminals. Note that this process also generates a sentence x , which is obtained by concatenating every terminal symbol in z , meaning that:

$$P(z|\theta) = P(x, z|\theta). \quad (2.5)$$

Some questions arise when applying this model to real parsing applications like grammar induction:

Parsing Given a PCFG G , how to find the best (highest probability) parse among all possible parses?

Learning Where do the probabilities, or rule weights θ come from?

The nice property of PCFGs is that there is a very general solution for these questions. We first discuss the first question in Section 2.3.2, and then deal with the second question in Section 2.3.3.

2.3.2 CKY Algorithm

Let us define some notations first. We assume the input sentence is a length n sentence, $x = x_1x_2 \cdots x_n$ where $x_i \in \Sigma$. For $i \leq j$, $x_{i,j} = x_ix_{i+1} \cdots x_j$ denotes an input substring. We assume the grammar is in CNF (Section 2.1.1), which makes the discussion much simpler.

Given an input sentence x and a PCFG with parameters θ , the goal of parsing is to solve the following argmax problem:

$$z' = \arg \max_{z \in \mathcal{Z}(x)} P(z|\theta), \quad (2.6)$$

where $\mathcal{Z}(x)$ is a set of all possible parses on x . Now we describe a general algorithm to solve this problem in polynomial time called the CKY algorithm, which also plays an essential role in parameter estimation of θ discussed in Section 2.3.3.

For the moment we simplify the problem as calculating the *probability* of the best parse instead of the best parse itself (Eq. 2.6). We later describe that the argmax problem can be solved with a small modification to this algorithm.

The CKY algorithm is a kind of chart parsing. For an input string, there are too many, exponential number of parses, which prohibit to enumerate one by one. To enable search in this large space, a chart parser divides the problem into subproblems, each of which analyze a small span $x_{i,j}$ and then is combined into an analysis of a larger span.

Let C be a chart, which gives mapping from a signature of a subspan, or an item (i, j, N) to a real value. That is, each cell of C keeps the probability of the best (highest score) analysis for a span $x_{i,j}$ with symbol N as its root. Algorithm 1 describes the CKY parsing, in which each chart cell is filled recursively. The procedure $\text{FILL}(i, j, N)$ return a value with memoization. This procedure is slightly different from the ones found in some textbooks (Manning and Schütze, 1999), which instead fill chart cells in specific order. We do not take this approach since our memoization-based technique need not care about correct order of filling chart cells, which is somewhat involved for more complex grammars such as the one we present in Chapter 5.

Algorithm 1 CKY Parsing

```

1: procedure PARSE( $x$ )
2:    $C[1, n, S] \leftarrow \text{FILL}(1, n, S)$  ▷ Recursively fills chart cells.
3:   return  $C[1, n, S]$ 
4: end procedure
5: procedure FILL( $i, j, N$ )
6:   if  $(i, j, N) \notin C$  then
7:     if  $i = j$  then
8:        $C[i, i, N] \leftarrow \theta_{N \rightarrow x_i}$  ▷ Terminal expansion.
9:     else
10:       $C[i, j, N] \leftarrow \max_{N \rightarrow A \ B \in R; k \in [i, j]} \theta_{N \rightarrow A \ B} \times \text{FILL}(i, k, A) \times \text{FILL}(k, j, B)$ 
11:    end if
12:  end if
13:  return  $C[i, j, N]$ 
14: end procedure

```

The crucial point in this algorithm is the recursive equation in line 10. The assumption here is that since the grammar is context-free, the parses of subspans (e.g., spans with signatures (i, k, A) and (k, j, B)) in the best parse of $x_{i,j}$ should also be the best parses at subspan levels. The goal of the algorithm is to fill the chart cell for an item $(1, n, S)$ where S is the start symbol, which corresponds the analysis of the full span (whole sentence).

To get the best parse, what we need to modify in the algorithm 1 is to keep backpointers to the cells of the best children when filling each chart cell during lines 7–11. This is commonly done by preparing another chart, in which each cell keeps not a numerical value but backpointers into other cells in that chart. The resulting algorithm is called the *Viterbi* algorithm, the details of which are found in the standard textbooks (Manning and Schütze, 1999).

The behavior of the CKY parsing is characterized by the procedure of filling chart cells in lines 7–11 of Algorithm 1. We often write this procedure visually as in Figure 2.22. With this specification we can analyze the time complexity of the CKY algorithm, which is $O(n^3|P|)$ where $|P|$ is the number of allowed rules in the grammar because each rule in Figure 2.22 is performed only once¹⁰ and there are at most $O(n^3|P|)$ ways of instantiations for BINARY rules.¹¹

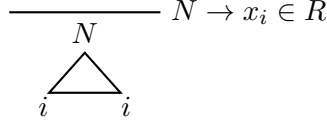
2.3.3 Learning parameters with EM algorithm

Next we briefly describe how rule weights θ can be estimated given a collection of input sentences. This is the setting of *unsupervised* learning. In supervised learning, we can often learn parameters more easily by counting rule occurrences in the training treebank (Collins, 1997; Johnson, 1998b; Klein and Manning, 2003).

¹⁰Once the chart cell of some item is calculated, we can access to the value of that cell in $O(1)$.

¹¹TERMINAL rules are instantiated at most $O(n|N|)$ ways, which is smaller than $O(n^3|R|)$.

TERMINAL:



BINARY:

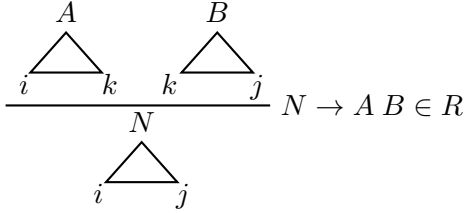


Figure 2.22: Inference rules of the CKY algorithm. **TERMINAL** rules correspond to the terminal expansion in line 8 of the Algorithm 1; **BINARY** rules correspond to the one in line 10. Each rule specifies how an analysis of a larger span (below —) is derived from the analyses of smaller spans (above —) provided that the input and grammar satisfy the side conditions in the right of —.

EM algorithm Assume we have a set of training examples $\mathbf{x} = x^{(1)}, x^{(2)}, \dots, x^{(m)}$. Each example is a sentence $x^{(i)} = x_1^{(i)} x_2^{(i)} \dots x_{n_i}^{(i)}$ where n_i is the length of the i -th sentence. Our goal is to estimate parameters θ given \mathbf{x} , which is good in some criteria.

The EM algorithm is closely related to maximum likelihood estimation in that it tries to estimate θ , which maximizes the following log-likelihood of the observed data \mathbf{x} :

$$L(\theta, \mathbf{x}) = \sum_{1 \leq i \leq m} \log p(x^{(i)} | \theta) = \sum_{1 \leq i \leq m} \log \sum_{z \in \mathcal{Z}(x^{(i)})} p(x^{(i)}, z | \theta), \quad (2.7)$$

where $p(x^{(i)}, z | \theta)$ is given by Eq. 2.4 due to Eq. 2.5. However, calculating θ that maximizes this objective is generally intractable (Dempster et al., 1977). The idea of EM algorithm is instead of getting the optimal θ , trying to increase Eq. 2.7 up to some point to find the locally optimal values of θ , starting from some initial values θ_0 . It is an iterative procedure and updates parameters as $\theta^{(0)} \rightarrow \theta^{(1)} \rightarrow \dots$ until specific number of iterations (or until $L(\theta, \mathbf{x})$ does not increase).

Each iteration of the EM algorithm proceeds as follows.

E-step Given the current parameters $\theta^{(t)}$, calculate the expected counts of each rule $e(r | \theta^{(t)})$ as

$$e(r | \theta^{(t)}) = \sum_{1 \leq i \leq m} e_{x^{(i)}}(r | \theta^{(t)}), \quad (2.8)$$

where $e_x(r | \theta^{(t)})$ is the expected counts of r in a sentence x , given by

$$e_x(r | \theta^{(t)}) = \sum_{z \in \mathcal{Z}(x)} p(z | x) f(r, z). \quad (2.9)$$

where $f(r, z)$ is the number of times that r appears in z . As in the parsing problem in Eq. 2.6,

it is impossible to directly calculate Eq. 2.9 by enumerating every parse. Below, we describe how this calculation becomes possible with the dynamic programming algorithm called the inside-outside algorithm, which is similar to CKY.

M-step Update the parameters as follows:

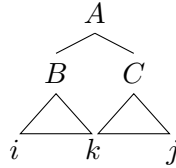
$$\theta_{A \rightarrow \beta}^{(t+1)} = \frac{e(A \rightarrow \beta | \theta^{(t)})}{\sum_{\alpha: A \rightarrow \alpha \in R} e(A \rightarrow \alpha | \theta^{(t)})} \quad (2.10)$$

This update is similar to the standard maximum likelihood estimation in the supervised learning setting, in which we observe the explicit counts of each rule. In the EM algorithm we do not explicitly observe rule counts, so we use the expected counts calculated with the previously estimated parameters. We can show that this procedure always increases the log likelihood (Eq. 2.7) until convergence though the final parameters are not globally optimum.

Inside-outside algorithm We now explain how the expected rule counts $e_x(r|\theta)$ are obtained for each r given sentence x . Let r be a binary rule $r = A \rightarrow B C$. First, it is useful to decompose $e_x(r|\theta)$ as the expected counts on each subspan as follows:

$$e_x(r|\theta) = \sum_{1 \leq i \leq k \leq j \leq n_x} e_x(z_{i,k,j,r}|\theta). \quad (2.11)$$

$e_x(z_{i,k,j,r}|\theta)$ is the expected counts of an event that the following fragment occurs in a parse z .



$z_{i,k,j,r}$ is an indicator variable¹² that is 1 if the parse contains the fragment above. Because the expected counts for an indicator variable are the same as the conditional probability of that variable (Bishop, 2006), we can rewrite $e_x(z_{i,k,j,r}|\theta)$ as follows:

$$e_x(z_{i,k,j,r}|\theta) = p(z_{i,k,j,r} = 1 | x, \theta) \quad (2.12)$$

$$= \frac{p(z_{i,k,j,r} = 1, x | \theta)}{p(x | \theta)}. \quad (2.13)$$

Intuitively the numerator in Eq. 2.13 is the total probability for generating parse trees that yield x and contain the fragment $z_{i,k,j,r}$. The denominator $p(x|\theta)$ is the marginal probability of the sentence x .

We first consider how to calculate $p(x|\theta)$, which can be done with a kind of CKY parsing; what we have to modify is just to replace the max operation in the line 10 in Algorithm 1 by the

¹²We omit dependence for x for simplicity.

summation operation. Then each chart cell $C[i, j, N]$ keeps the marginal probability for a subspan $x_{i,j}$ rooted at N . After filling the chart, $C[1, n_x, S]$ is the sentence marginal probability $p(x|\theta)$. The marginal probability for the signature (i, j, N) is called the *inside* probability, and this chart algorithm is called the inside algorithm, which calculates inside probabilities by filling chart cells recursively.

Calculation of $p(z_{i,k,j,r} = 1, x|\theta)$ is more elaborate so we only sketch the idea here. Analogous to the inside probability introduced above, we can also define the outside probability $O(i, j, N)$, which is the marginal probability for the outside of the span with signature (i, j, N) ; that is,

$$O(i, j, N) = p(x_{1,i-1}, N, x_{j+1,n}|\theta), \quad (2.14)$$

in which N roots the subspan $x_{i,j}$. Since $I(i, j, N) = p(x_{i,j}|N, \theta)$, given $r = N \rightarrow A B$ we obtain:

$$p(z_{i,k,j,r} = 1, x_{i,j}|N, \theta) = \theta_{N \rightarrow A B} \times I(i, k, A) \times I(k, j, B), \quad (2.15)$$

which is the total probability of parse trees for the subspan $x_{i,j}$ that contains the fragment indicated by $z_{i,k,j,r}$. Combining these two terms, we obtain:

$$p(z_{i,k,j,r} = 1, x|\theta) = p(z_{i,k,j,r} = 1, x_{i,j}|N, \theta) \times p(x_{1,i-1}, N, x_{j+1,n}|\theta) \quad (2.16)$$

$$= \theta_{N \rightarrow A B} \times I(i, k, A) \times I(k, j, B) \times O(i, j, N). \quad (2.17)$$

2.3.4 When the algorithms work?

So far we have assumed the underlying grammar is a PCFG, for which we have introduced two algorithms, the CKY algorithm and the EM algorithm with inside-outside calculation. However as we noted in the beginning of Section 2.3.1, the scope of these algorithms is not limited to PCFGs. What is the precise condition under which these algorithms can be applied?

PCFGs are a special instance of weighted CFGs, in which each rule has a weight but the sum of rule weights from a parent nonterminal (Eq. 2.2) is not necessarily normalized. As we see next, the split bilexical grammars in Section 2.3.5 can always be converted to a weighted CFG but may not be converted to a PCFG.

The CKY algorithm can be applied to *any* weighted CFGs. This is easily verified because only the assumption in the Algorithm 1 is that the grammar is context-free for being able to divide a larger problem into smaller subproblems.

The condition for the inside-outside algorithm is more involved. Let us assume that we have a generative model of a parse, which is not originally parameterized with a PCFG, and also we have a weighted CFG designed so that the score that this weighted CFG gives to a (CFG) parse is the same as the probability that the original generative model assigns to the corresponding parse in the original form (not CFG) (The model in Section 2.3.5 is an example of such cases.).

Then, the necessary condition on this weighted CFG is that there is no spurious ambiguity between two representations (Section 2.1.4); that is, a CFG parse can be uniquely converted to the parse in the original form, and vice versa. The main reason why the spurious ambiguity cause a problem is that the quantities used to calculate expected counts (Eq. 2.13) are not correctly defined if the spurious ambiguity exists. For example, the sentence marginal probability $p(x|\theta)$ would not

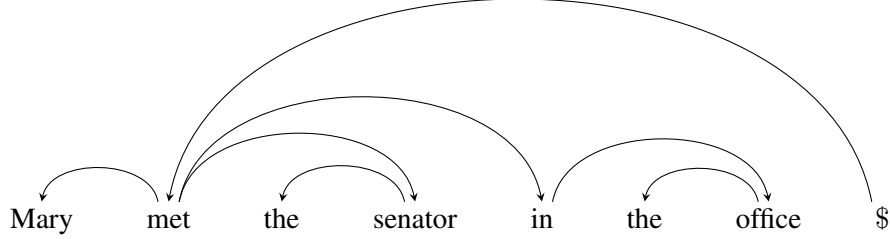


Figure 2.23: Example of a projective dependency tree generated by a SBG. \$ is always placed at the end of a sentence, which has only one dependent in the left direction.

be equal to the inside probability for the whole sentence $I(1, n_x, S)$ when the spurious ambiguity exists since if there are multiple CFG derivations for a single parse in the original form the inside probability calculated with the weighted CFG would overestimate the true sentence marginal probability. The same issue happens for another quantity in Eq. 2.17. On the other hand, if there is one-to-one correspondence between two representations, it should hold at the smaller subspan levels and it is this transparency that guarantees the correctness of the inside-outside algorithm even if the grammar is not strictly a PCFG.

2.3.5 Split bilexical grammars

The split bilexical grammars, or SBGs (Eisner, 2000) is a notationally simpler variant of split head-automaton grammars (Eisner and Satta, 1999). Here we describe this model as a generalization of the specific model described in Section 2.3.6, the dependency model with valence (DMV) (Klein and Manning, 2004). We will give a somewhat in-depth explanation of this grammar below because it will be the basis of our proposal in Chapter 5.

The explanation below basically follows Eisner and Smith (2010). A SBG defines a distribution over projective dependency trees. This model can easily be converted to an equivalent weighted CFG, although some effort is needed to remove the *spurious ambiguity*. We will show that by removing it the time complexity can also be improved from $O(n^5)$ to $O(n^3)$. In Chapter 5 we will invent the similar technique for our model that follows the left-corner parsing strategy.

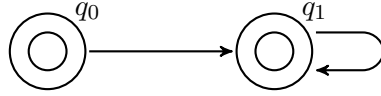
Model A (probabilistic) SBG is a tuple $G_{SBG} = (\Sigma, \$, L, R)$. Σ is an alphabet of words that may appear in a sentence. $\$ \notin \Sigma$ is a distinguished root symbol, which we describe later; let $\bar{\Sigma} = \Sigma \cup \{\$\}$. L and R are functions from $\bar{\Sigma}$ to probabilistic ϵ -free finite-state automata over $\bar{\Sigma}$; that is, for each $a \in \bar{\Sigma}$ the SBG specifies “left” and “right” probabilistic FSAs, L_a and R_a . We write $q \xrightarrow{a'} r \in R_a$ to denote a state transition from q to r by adding a' to a ’s right dependents when the current right state of a is q . Also each model defines $init(L_a)$ and $init(R_a)$ that return the set of initial states for a in either direction (usually the initial state is unique given the head a and the direction). $final(L_a)$ is a set of final states; $q \in final(L_a)$ means that a can stop generating its

left dependents. We will show that by changing the definitions of these functions several generative models over dependency trees can be represented in a unified framework.

The model generates a sentence $x_1 x_2 \cdots x_n \$$ along with a parse, given the root symbol $\$$, which is always placed at the end. An example of a parse is shown in Figure 2.23. SBGs define the following generative process over dependency trees:

1. The root symbol $\$$ generates a left dependent a from $q \in \text{init}(L_\$)$. a is regarded as the conventional root word in a sentence (e.g., *met* in Figure 2.23).
2. Recursively generates a parse tree. Given the current head a , the model generates its left dependents and its right dependents. This process is head-outward, meaning that the closest dependent is generated first. For example, the initial left state of *met*, $q \in \text{init}(L_{\text{met}})$ generates *Mary*. The right dependents are generated as follows: First *senator* is generated from $q_0 \in \text{init}(R_{\text{met}})$. Then the state may be changed by a transition $q_0 \xrightarrow{\text{senator}} q_1$ to $q_1 \in R_{\text{met}}$, which generates *in*. The process stops when every token stops generating its left and right dependents.

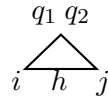
This model can generalize several generative models over dependency trees. Given a , L_a and R_a define distributions over a 's left and right children, respectively. Since the automata L_a and R_a have the current state, we can define several distributions by customizing the topology of state transitions. For example if we define the automata of the form:



it would allow the first (closest) dependent to be chosen differently from the rest (q_0 defines the probability of the first dependent). If we remove q_0 , the resulting automata are $\odot \rightleftarrows$ with a single state q_1 , so token a 's left (or right) dependents are conditionally independent of one another given a .¹³

Spurious ambiguity in naive $O(n^5)$ parsing algorithm We now describe that a SBG can be converted to a weighted CFG, though the distributions associated with L_a and R_a cannot always be encoded in a form of a PCFG. Also, as we see below, the grammar suffers from the spurious ambiguity, which prevent us to apply the inside-outside algorithm (see Section 2.3.4).

The key observation for this conversion is that we can represent a subtree of SBGs as the following triangle, which can be seen as a special case of a subtree used in the ordinary CKY parsing.



¹³SBGs can also be used to encode second-order *adjacent* dependencies, i.e., a 's left or right dependent to be dependent on its sibling word just generated before, although in this case there exists more efficient factorization that leads to a better asymptotic runtime (McDonald and Pereira, 2006; Johnson, 2007).

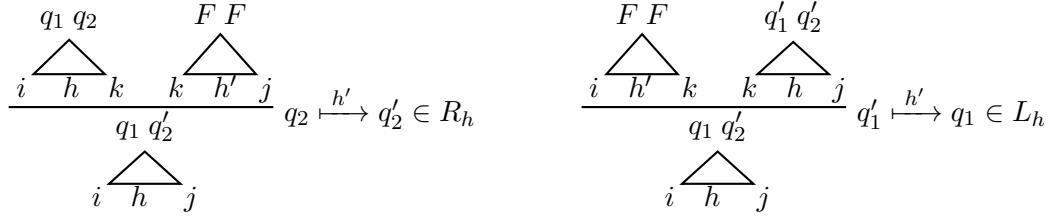


Figure 2.24: Binary inference rules in the naive CFG conversion. F means the state is a final state in that direction. Both left and right consequent items (below ---) have the same item but from different derivations, suggesting 1) the weighted grammar is not a PCFG; and 2) there is the spurious ambiguity.

The main difference from the ordinary subtree representation is that it is decorated with an additional index h , which is the position of the head word. For example in the analysis of Figure 2.23, a subtree on *met the senator* is represented by setting $i = 2, j = 4, h = 2$. q_1 and q_2 are the current h 's left and right states respectively.

We can assume a tuple (h, q_1, q_2) to comprise a nonterminal symbol of a CFG. Then the grammar is a PCFG if the normalization condition (Eq. 2.2) is satisfied for every such tuple. Note now each rule looks like $(h, q_1, q_2) \rightarrow \beta$.

Figure 2.24 explains why the grammar cannot be a PCFG. We can naturally associate PCFG rule weights for these rules with transition probabilities given by the automata L_h and R_h .¹⁴ However, then the sum of rule weights of the converted CFG starting from symbol (a, q_1, q'_2) is not equal to 1. The left rule of Figure 2.24 means the converted CFG would have rules of the form $(a, q_1, q'_2) \rightarrow (a, q_1, q_2) (a', F, F)$. The weights associated with these rules are normalized across $a' \in \Sigma$, as state transitions are deterministic given q'_2 and a' . The problem is that the same signature, i.e., (a, q_1, q'_2) on the same span can also be derived from another rule in the right side of Figure 2.24. This distribution is also normalized, meaning that the sum of rule weights is not 1.0 but 2.0, and thus the grammar is a weighted CFG, not a PCFG. The above result also suggests that the grammar suffers from the spurious ambiguity. Below we describe how this ambiguity can be removed with modifications.

As a final note, the time complexity of the algorithm in Figure 2.24 is very inefficient, $O(n^5)$ because there are five free indexes in each rule. This is in contrast to the complexity of original CKY parsing, which is $O(n^3)$. The refinement described next also fix this problem, and we obtain the $O(n^3)$ algorithm for parsing general SBGs.

$O(n^3)$ algorithm with head-splitting The main reason why the algorithm in Figure 2.24 causes the spurious ambiguity is because given a head there is no restriction in the order of collecting its left and right dependents. This problem can be handled by introducing new items called *half constituents* denoted by \triangleleft (*left constituents*) and \triangleright (*right constituents*), which represent left and right span separately given a head. For example in the dependency tree in Figure 2.23, a phrase

¹⁴Here and the following, we occasionally abuse the notation and use L_h or R_h to mean the automaton associated with word x_h at index position h .

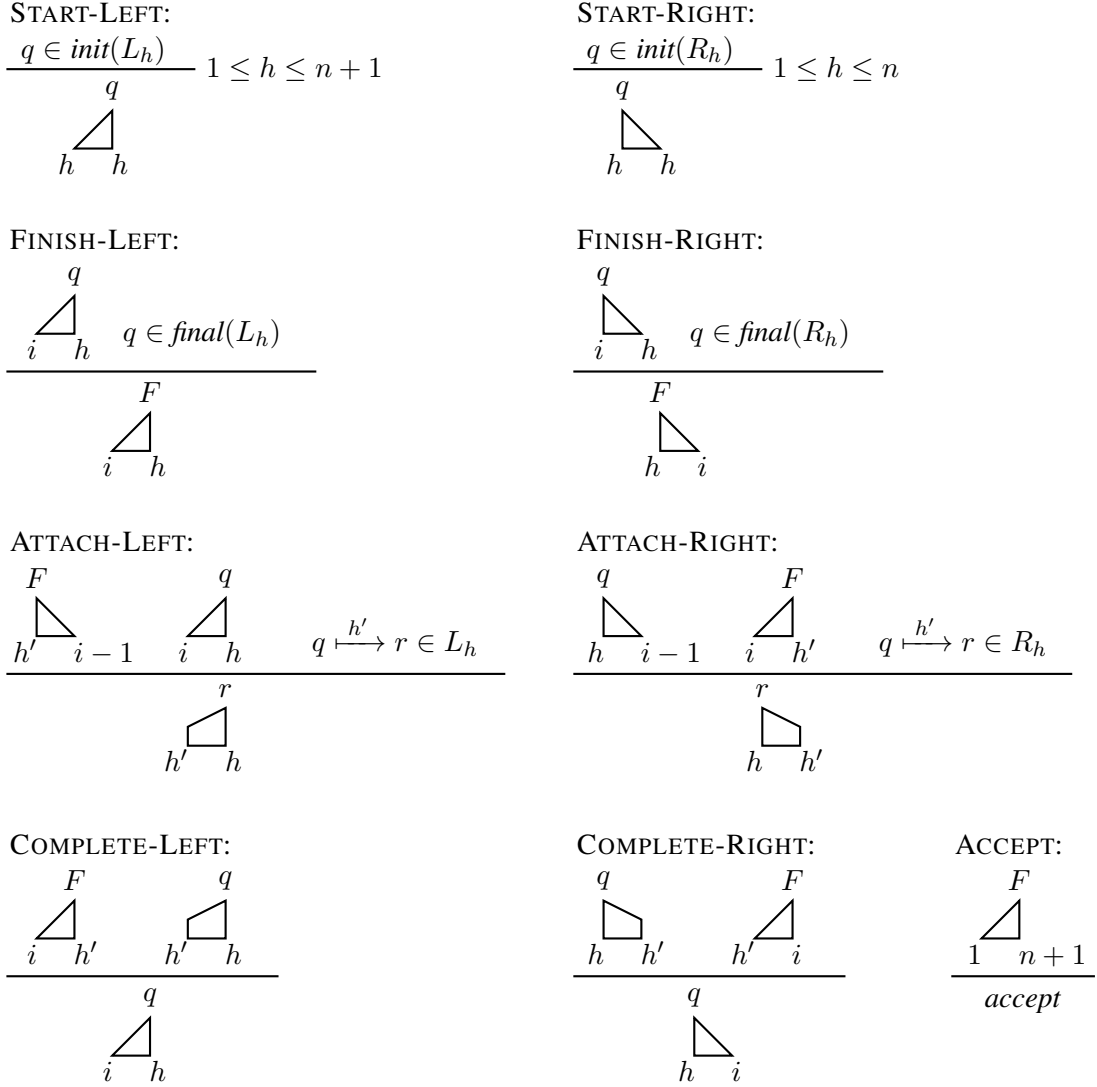






Figure 2.25: An algorithm for parsing SBGs in $O(n^3)$ given a length n sentence. The $n+1$ -th token is a dummy root token $\$,$ which only has one left dependent (sentence root). $i, j, h,$ and h' are index of a token in the given sentence while $q, r,$ and F are states. L_h and R_h are left and right FSA of the h -th token in the sentence. Each item as well as a statement about a state (e.g., $r \in \text{final}(L_p)$) has a weight and the weight of a consequent item (below ---) is obtained by the product of the weights of its antecedent items (above ---).

“Mary met” comprises a left constituent while “met the senator” comprises a right constituent. In the new algorithm these two constituents are expanded separately with each other and there is no *mixed* states (q_1, q_2) as in the items in Figure 2.24. Eliminating these mixed states is the key to eliminate the spurious ambiguity.

Figure 2.25 shows new algorithm, which can be understood as follows:

- ATTACH-LEFT and COMPLETE-LEFT (or the RIGHT counterpart) are the essential components of the algorithm. The idea is when combining two constituents headed by h' and h ($h' < h$) into a large constituent headed by h , we decompose an original constituent  into its left and right half constituents, and combine those fragments in order. ATTACH-LEFT does the first part, i.e., collects the right constituent . The resulting trapezoid  represents an intermediate parsing state, which means the recognition of the right half part of h' has finished while the remaining left part yet unfinished. COMPLETE-LEFT does the second part and collects the remaining left constituent . ATTACH-RIGHT and COMPLETE-RIGHT do the opposite operations and collect the right dependents of some head.
- On this process, the state F in both left and right constituents ensure that they can be a dependent of others.
- START-LEFT and START-RIGHT correspond to the terminal rules of the ordinary CKY algorithm (Figure 2.22) though we segment it into the left and right parts. Note that the root symbol $\$$ at the $n + 1$ position only applies START-LEFT because it must not have any right dependents. Commonly the left automaton $L_\$$ is designed to have only one dependent; otherwise, the algorithm may allow the fragmental parses with more than one root tokens.
- Differently from the inference rules in Figure 2.24, we put the state transitions, e.g., $q \xrightarrow{h'} r \in R_h$ as antecedent items (above \rightarrow) of each rule instead of the side condition. These modifications are to make the weight calculation at each rule more explicit. Specifically, when we develop a model in this framework, each state transition, i.e., $q \in \text{init}(L_h)$, $q \in \text{final}(L_h)$, and $q \xrightarrow{h'} r \in L_h$ (or for R_h) has an associated weight. Also when we run the CKY or the related algorithm, each chart cell that corresponds to some constituent (triangle or trapezoid) has a weight. Thus, this formulation makes explicit that the weight of the consequent item (below \rightarrow) is obtained by the product of all weights of the antecedent items (above \rightarrow). We describe the particular parameterization of these transitions to achieve DMV in Section 2.3.6.
- The grammar is not in CNF since it contains unary rules at internal positions. The inside-outside algorithm can still be applied by assuming null element (which has weight 1) in either child position in Algorithm 1.


There is no spurious ambiguity. However, again this grammar is not always a PCFG. In particular, the grammar for a dependency model with valence (DMV), which we describe next, is not a

| Transition | Weight (DMV parameters) |
|-----------------------------------|---|
| $q_0 \in \text{init}(L_h)$ | 1.0 |
| $q_0 \in \text{final}(L_h)$ | $\theta_s(\text{STOP} h, \leftarrow, \text{TRUE})$ |
| $q_1 \in \text{final}(L_h)$ | $\theta_s(\text{STOP} h, \leftarrow, \text{FALSE})$ |
| $q_0 \xrightarrow{d} q_1 \in L_h$ | $\theta_A(d h, \leftarrow) \cdot \theta_s(\neg\text{STOP} h, \leftarrow, \text{TRUE})$ |
| $q_1 \xrightarrow{d} q_1 \in L_h$ | $\theta_A(d h, \leftarrow) \cdot \theta_s(\neg\text{STOP} h, \leftarrow, \text{FALSE})$ |

Figure 2.26: Mappings between FSA transitions of SBGs and the weights to achieve DMV. θ_s and θ_a are parameters of DMV described in the body. The right cases (e.g., $q_0 \in \text{init}(R_a)$) are omitted but defined similarly. h and d are both word types, not indexes in a sentence (contrary to Figure 2.25).

PCFG. See the FINISH-LEFT rule in the algorithm. A particular model such as DMV may associate a score for this rule to explicitly model an event that a head h stops generating its left dependents. In such cases, the weights for CFG rules $(F, h) \rightarrow (q, h)$ do not define a correct (normalized) distribution given the parent symbol (F, h) . This type of inconsistency happens due to discrepancy between the underlying parsing strategies in two representations: The PCFGs assume the tree generation is a top-down process while the SBGs assume it is bottom-up. Nevertheless we can use the inside-outside algorithm as in PCFGs because there is no spurious ambiguity and each derivation in a CFG parse correctly gives a probability that the original SBG would give to the corresponding dependency tree.

Also time complexity is improved to $O(n^3)$. This is easily verified since there appear at most three indexes on each rule. The reason of this reduction is we no longer use full constituents with

a head index  h , which itself consumes three indexes, leading to an asymptotically inefficient algorithm.

2.3.6 Dependency model with valence

Now it is not so hard to formulate the famous model, dependency model with valence (DMV), on which our unsupervised model is based, as a special instance of SBGs. This can be done by defining transitions of each automaton as well as the associated weights. In DMV, each L_h or R_h given head h has only two states q_0 and q_1 , both of which are in finished states, i.e., FINISH-LEFT and FINISH-RIGHT in Figure 2.25 can always be applied. q_0 is the initial state and $q_0 \xrightarrow{h} q_1$ while $q_1 \xrightarrow{h} q_1$, meaning that we only distinguish the generation process of the first dependent from others.

The associated weights for transitions in Figure 2.25 are summarized in Figure 2.26. Each weight is a product of DMV parameters, which are classified into two types of multinomial distributions θ_s and θ_A . Generally we write $\theta_{\text{TYPE}}(d|c)$ for a multinomial parameter in which TYPE defines a type of multinomial, c is a conditioning context, and d is a decision given the context. DMV has the following two types of parameters:

- $\theta_s(\text{stop}|h, \text{dir}, \text{adj})$: A Bernoulli random variable to decide whether or not to attach further dependents in the current direction $\text{dir} \in \{\leftarrow, \rightarrow\}$. The decision $\text{stop} \in \{\text{STOP}, \neg\text{STOP}\}$. The

adjacency $adj \in \{\text{TRUE}, \text{FALSE}\}$ is the key factor to distinguish the distributions of the first and other dependents. It is TRUE if h has no dependent yet in dir direction.

- $\theta_A(d|h, dir)$: A probability that d is attached as a new dependent of h in dir direction.

The key behind the success of the DMV was the introduction of the valence factor in stop probabilities (Klein and Manning, 2004). Intuitively, this factor can capture the difference of the expected number of dependents for each head. For example, in English, a verb typically takes one dependent (subject) in the left direction while several dependents in the right direction. DMV may capture this difference with a higher value of $\theta_s(\neg\text{STOP}|h, \leftarrow, \text{TRUE})$ and a lower value of $\theta_s(\neg\text{STOP}|h, \leftarrow, \text{FALSE})$. On the other hand, in the right direction, $\theta_s(\neg\text{STOP}|h, \rightarrow, \text{FALSE})$ might be higher, facilitating to attach several dependents.

Inference With the EM algorithm, we try to update parameters θ_s and θ_A . This is basically done with the inside-outside algorithm though one complicated point is that some transitions in Figure 2.26 are associated with products of parameters, not a single parameter. This situation contrasts with the original inside-outside algorithm for PCFGs where each rule is associated with only a single parameter (e.g., $A \rightarrow \beta$ and $\theta_{A \rightarrow \beta}$).

In this case the update can be done by first collecting the expected counts of each transition in a SBG, and then converting it to the expected counts of a DMV parameter. For example, let $e_x(\text{ATTACH-LEFT}, q, h, d|\theta)$ be the expected counts of the ATTACH-LEFT rule between head h with state q and dependent h' in a sentence x . We can obtain $e_x(h, d, \leftarrow|\theta)$, the expected counts of an attachment parameter of DMV as follows:

$$e_x(h, d, \leftarrow|\theta) = e_x(\text{ATTACH-LEFT}, q_0, h, d|\theta) + e_x(\text{ATTACH-LEFT}, q_1, h, d|\theta). \quad (2.18)$$

These are then normalized to update the parameters (as in Section 2.3.3). Similarly the counts of the non-stop decision $e_x(h, \neg\text{STOP}, \leftarrow, \text{TRUE}|\theta)$, associated with $\theta_s(\neg\text{STOP}|h, \leftarrow, \text{TRUE})$, is obtained by:

$$e_x(h, \neg\text{STOP}, \leftarrow, \text{TRUE}|\theta) = \sum_{h'} e_x(\text{ATTACH-LEFT}, q_0, h, h'|\theta), \quad (2.19)$$

where h' is a possible left dependent (word type) of h .

2.3.7 Log-linear parameterization

In Chapter 5, we build our model based on an extended model of DMV with *features*, which we describe in this section. We call this model *featurized DMV*, which first appeared in Berg-Kirkpatrick et al. (2010). We use this model since it is relatively a simple extension to DMV (among others) while known to boost the performance well.

The basic idea is that we replace each parameter of the DMV as the following log-linear model:

$$\theta_A(d|h, dir) = \frac{\exp(\mathbf{w}^\top \mathbf{f}(d, h, dir, A))}{\sum_{d'} \exp(\mathbf{w}^\top \mathbf{f}(d', h, dir, A))}, \quad (2.20)$$

where \mathbf{w} is a weight vector and $\mathbf{f}(d, h, \text{dir}, A)$ is a feature vector for an event that h takes d as a dependent in dir direction. Note that contrary to the more familiar log-linear models in NLP, such as the conditional random fields (Lafferty et al., 2001; Finkel et al., 2008), it does not try to model the whole structure with a single log-linear model. Such approaches make it possible to exploit more richer global structural features though inference gets more complex and challenging in particular in an unsupervised setting (Smith and Eisner, 2005; Ammar et al., 2014).

In this model, the features can only be exploited from the conditioning context and decision of each original DMV parameter. The typical information captured with this method is the back-off structures between parameters. For example, some feature in \mathbf{f} is the one ignoring direction, which facilitates sharing of statistical strength of attachments between h and d . Berg-Kirkpatrick et al. (2010) also report that adding back-off features that use the coarse POS tags is effective, e.g., ones replacing actual h or d with a coarse indicator, such as whether h belongs to a (coarse) noun category or not, when the original dataset provides finer POS tags (e.g., pronoun or proper noun).

The EM-like procedure can be applied to this model with a little modification, which instead of optimizing parameters θ directly, optimizes weight vector \mathbf{w} . The E-step is exactly the same as the original algorithm. In the M-step, we optimize \mathbf{w} to increase the marginal log-likelihood (Eq. 2.7) using the gradient-based optimization method such as L-BFGS (Liu and Nocedal, 1989) with the expected counts obtained from the E-step. In practice, we optimize the objective with the regularization term to prevent overfitting.

2.4 Previous Approaches in Unsupervised Grammar Induction

This section summarizes what has been done in the study of unsupervised grammar induction in particular in this decade from Klein and Manning (2004), which was the first study breaking the simple baseline method in English experiments. Here we focus on the setting of *monolingual* unsupervised parsing, which we first define in Section 2.4.1. Related approaches utilizing some kind of supervised information, such as semi-supervised learning (Haghighi and Klein, 2006) or transfer learning in which existing high quality parsers (or treebanks) for some languages (typically English) are transferred into parsing models of other languages (McDonald et al., 2011; Naseem et al., 2012; McDonald et al., 2013; Täckström et al., 2013) exist. These approaches typically achieve higher accuracies though we do not touch here.

2.4.1 Task setting

The typical setting of unsupervised grammar induction is summarized as follows:

- During training, the model learns its parameters using (unannotated) sentences only. Sometimes the model uses external resources, such as Wikipedia articles (Mareček and Straka, 2013) to exploit some statistics (e.g., n-gram) in large corpora but does not rely on any syntactic annotations.

- To remedy the data sparseness (or the learning difficulty), often the model assumes part-of-speech (POS) tags as the input instead of surface forms.¹⁵ This assumption greatly simplifies the problem though it loses much crucial information for disambiguation. For example, the model may not be able to disambiguate prepositional phrase (PP) attachments based on semantic cues as *supervised* parsers would do. Consider two phrases *eat sushi with tuna* and *eat sushi with chopsticks*. The syntactic structures for these two are different, but POS-based models may not distinguish between them as they both look the same under the model, e.g., VERB NOUN ADP NOUN; ADP is an adposition. Therefore the main challenge of unsupervised grammar induction is often to acquire more basic structures or the word order, such that an adjective tends to modify a noun.
- The POS-based models are further divided into two categories, whether it can or cannot access to the *semantics* of each POS tag. The example of the former is Naseem et al. (2010), which utilizes the information, e.g., a verb tend to be the root of a sentence. This approach is sometimes called *lightly* supervised learning. The latter approach, which we call *purely* unsupervised learning, does not access to such knowledge. In this case, the only necessary input for the model is the clustering of words, not the *label* for each cluster. This is advantageous in practice since it can be based on the output of some unsupervised POS tagger, which cannot identify the semantics (label) of each induced cluster. Though the problem settings are slightly different in two approaches, we discuss both here as it is unknown what kind of prior linguistic knowledge is necessary for learning grammars. Note that it is also an ongoing study how to achieve lightly supervised learning from the output of unsupervised POS taggers with a small amount of manual efforts (Bisk et al., 2015).

Evaluation The evaluation of unsupervised systems is generally difficult and controversial. This is particularly true in unsupervised grammar induction. The common procedure, which most works described below employ, is to compare the system outputs and the gold annotated trees just as in the supervised case. That is, we evaluate the quality of the system in terms of accuracy measure, which is precision, recall, and F1-score for constituent structures and an attachment score for dependency structures. This is inherently flawed in some sense mainly because it cannot take into account the variation in the notion of linguistically *correct* structures. For example, some dependency structures, such as coordination structures, are analyzed in several ways (see also Section 3.1); each of which is *correct* under a particular syntactic theory (Popel et al., 2013) but the current evaluation metric penalizes unless the prediction of the model matches the gold data currently used. We do not discuss the solution to this problem here. However, we try to minimize the effect of such variations in our experiments in Chapter 5. See Section 5.3.3 for details.

2.4.2 Constituent structure induction

As we saw in Section 2.3.3, the EM algorithm provides an easy way for learning parameters of any PCFGs. This motivated the researchers to use the EM algorithm for obtaining syntactic trees without

¹⁵ Some work, e.g., Seginer (2007) does not assume this convention as we describe in Section 2.4.4.

human efforts (annotations). In early such attempts, the main focus for the induced structures has been phrase-structure trees.

However, it has been well known that such EM-based approaches perform poorly to recover the syntactic trees that linguists assume to be correct (Charniak, 1993; Manning and Schütze, 1999; de Marcken, 1999). The reasons are mainly two-folds: One is that the EM algorithm is just a hill climbing method so it cannot reach the global optimum solution. Since the search space of the grammar is highly complex, this local maxima problem is particularly a severe problem; Carroll et al. (1992) observed that randomly initialized EM algorithms always converge to different grammars, which all are far from the target grammar. Another crucial problem is the inherent difficulty in the induction of PCFGs. In the general setting, the fixed structure for the model is just the start symbol and observed terminal symbols. The problem is that although terminal symbols are the most informative source for learning, that information does not correctly propagate to the higher level in the tree since each nonterminal label here is just an abstract symbol (hidden categorical variable) and has less meaning. For example, when the model has a rule $y_1 \rightarrow y_2 y_3$ and y_2 and y_3 dominate some subtrees, y_1 dominates a larger constituent but its relevance to the yield (i.e., dominated terminal symbols) sharply diminishes.

For these reasons, so far the only successful PCFG-based constituent structure induction methods are by giving some amount of supervision, e.g., constraints on possible bracketing (Pereira and Schabes, 1992) and possible rewrite rules (Carroll et al., 1992). Johnson et al. (2007) reported that the situation does not change with the sampling-based Bayesian inference method.

Non PCFG-based constituent structure induction has been explored since early 2000s with some success. The common idea behind these approaches is not collapsing each span into the (less meaningful) nonterminal symbols. Clark (2001) and Klein and Manning (2002) are such attempts, in which the model tries to learn whether some yields (n-gram) comprises a constituent or not. All parameters are connected to terminal symbols and thus the problem in propagating information from the terminal symbols is alleviated. Ponvert et al. (2011) present a chunking-based heuristic method to improve the performance in this line of models. Seginer (2007) is another successful constituent induction system; We discuss his method in Section 2.4.4 as it has some relevance to our approach.

2.4.3 Dependency grammar induction

Due to the difficulty in PCFG-based constituent structure induction, most recent works in PCFG induction has focused on dependency as its underlying structures. The dependency model with valence (DMV) (Klein and Manning, 2004) that we introduced in Section 2.3.6 is the most successful approach in such dependency-based models. As we saw, this model can be represented as an instance of weighted CFGs and thus parameter estimation is possible with the EM algorithm. This makes an extension on both model and inference easier, and results in many extensions on DMV in a decade as we summarize below.

We divide those previous approaches in largely two categories in whether the model relies on light supervision on dependency rules or not (see also Section 2.4.1). The goal of every study introduced below can be seen to identify the necessary bias or supervision for an unsupervised parser to learn accurate grammars without explicitly annotated corpora.

Purely unsupervised approaches Generally, purely unsupervised methods perform worse than the other, lightly supervised approaches (Bisk and Hockenmaier, 2012).

We first mention that the success of most works discussed below including the original DMV in Klein and Manning (2004) rely on a heuristic initialization technique often called the harmonic initializer, which we describe in details in Section 5.3.6. Since the EM algorithm is the local search method, it suffers from the local optima problem, meaning that it is sensitive to the initialization. Intuitively the harmonic initializer initializes the parameters to favor shorter dependencies. Gimpel and Smith (2012) reports that DMV *without* this initialization performs very badly; the accuracy on English experiments (Wall Street Journal portion of the Penn treebank) significantly drops from 44.1 to 21.3. Most works cited below rely on this technique, but some does not; in that case, we mention it explicitly (e.g., Mareček and Žabokrtský (2012)).

Bayesian modeling and inference are popular approach for enhancing the probabilistic models. In dependency grammar induction, Cohen and Smith (2009), Headden III et al. (2009), and Blunsom and Cohn (2010) are examples of such approaches. Cohen and Smith (2009) extend the baseline DMV model with somewhat complex priors called shared logistic normal priors, which enable to tie parameters of related POS tags (e.g., subcategories of nouns) to behave similarly. This is conceptually similar to the feature-based log-linear model (Berg-Kirkpatrick et al., 2010) that we introduced in Section 2.3.7. They employ variational EM for the inference technique.

Headden III et al. (2009) develop carefully designed Bayesian generative models, which are also estimated with the variational EM. This model is one of the few examples of a *lexicalized* model, i.e., utilizing words (surface forms) in additional to POS tags. This is a *partially* lexicalized model, meaning that the words that appear less than 100 times in the training data is unlexicalized. Another technique introduced in this paper is random initialization with model selection; they report that the performance improves by running a few iteration of EM in thousands of randomly initialized models and then picking up one with the highest likelihood. However, this procedure is too expensive and the later works do not follow it.

Blunsom and Cohn (2010) is one of the current state-of-the-art methods in purely unsupervised approach. In the shared task at the Workshop on Inducing Linguistic Structure (WILS) (Gelling et al., 2012), it performs competitively to the lightly supervised CCG-based approach (Bisk and Hockenmaier, 2012) on average across 10 languages. The model is partially lexicalized as in Headden III et al. (2009). Though the basic model is an extended model of the DMV, they encode the model on Bayesian tree substitution grammars (Cohn et al., 2010), which enable to model larger tree fragments than the original DMV does.

Mareček and Žabokrtský (2012) and Mareček and Straka (2013) present methods that learn the grammars using some principle of dependencies, which they call the *reducibility* principle. They argue that phrases that will be the dependent of another token (head) are often *reducible*, meaning that the sentence without such phrases is probably still grammatical. They calculate the reducibility of each POS n-gram using large raw text corpus from Wikipedia articles and develop a model that biases highly reducible sequences to become dependents. Mareček and Straka (2013) encode the reducibility on DMV and find that their method is not sensitive to initialization. This approach nicely exploits the property of heads and dependents and they report the state-of-the-art scores on the datasets of CoNLL shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007a).

Another line of studies on several extensions or heuristics for improving DMV has been explored by Spitzkovsky and colleagues. For example, Spitzkovsky et al. (2010b) report that sometimes the Viterbi objective instead of the EM objective leads to the better model and Spitzkovsky et al. (2010a) present the heuristics that starts learning from the shorter sentences only and gradually increases the training sentences. Spitzkovsky et al. (2013) is their final method. While the reported results are impressive (very competitive to Mareček and Straka (2013)), the method, which tries to avoid the local optima with the combination of several heuristics, e.g., changing the objective, forgetting some previous counts, and changing the training examples, is quite complex and it makes difficult to analyze which component most contributes to attained improvements.

The important study for us is Smith and Eisner (2006), which explores a kind of *structural* bias to favor shorter dependencies. This becomes one of our baseline models in Chapter 5; See section 5.3.5 for more details. We argue however that the motivation in their experiment is slightly different from the other studies cited above and us. Along with a bias to favor shorter dependencies, they also investigate the technique called *structural annealing*, in which the strength of the imposed bias is gradually relaxed. Note here that by introducing such new techniques, the number of adjustable parameters (i.e., hyperparameters) increases. Smith and Eisner (2006) choose the best setting of those parameters based on the *supervised* model setting, in which the annotated development data is used for choosing the best model. This is however not the unsupervised learning setting. In our experiments in Chapter 5, we thus do not explore the annealing technique and just compare the effects of different structural biases, that is, the shorter dependency length bias and our proposing bias of limiting center-embedding.

Lightly supervised approaches Naseem et al. (2010) is the first model utilizing the light supervision on dependency rules. The rules are specified declaratively as dependencies between POS tags such as VERB \rightarrow NOUN or NOUN \rightarrow ADJ. Then the model parameters are learned with the *posterior regularization* technique (Ganchev et al., 2010), which biases the posterior distribution at each EM iteration to put more weights on those specified dependency rules. Naseem et al. (2010) design 13 universal rules in total, and show state-of-the-art scores across a number of languages. This is not purely unsupervised approach but it gives an important upper bound on the required knowledge to achieve reasonable accuracies on this task. For example, Bisk and Hockenmaier (2013) demonstrate that the competitive scores to them is obtainable with a smaller amount of supervision by casting the model on CCG.

Søgaard (2012) present a heuristic method that does *not* learn anything, but just build a parse tree deterministically; For example it always recognizes the left-most verb to be the root word of the sentence. Although this is extremely simple, Søgaard reports it beats many systems submitted in the WILS shared task (Gelling et al., 2012), suggesting that often such declarative dependency rules alone can capture the basic word order of the language.

Grave and Elhadad (2015) is the recently proposed strong system, which also relies on the declarative rules. Instead of the generative models as in most works cited above, they formulate their model in a discriminative clustering framework (Xu et al., 2005), with which the objective becomes convex and the optimality is satisfied. This system becomes our strong baseline in the experiments in Chapter 5. Note that their system relies on in total 12 rules between POS tags. We explore how the competitive model to this system can be achieved with our structural constraints as

well as a smaller amount of supervision.

2.4.4 Other approaches

There also exist some approaches that do not induce dependency nor constituent structures directly. Typically for evaluation reasons the induced structure is converted to either form.

Common cover link Seginer (2007) presents his own grammar formalism called the common cover link (CCL), which looks similar to the dependency structure but differs in many points. For example, in CCL, every link between words is fully connected at every prefix position in the sentence. His parser and learning algorithm are fully incremental; He argues that the CCL structure as well as the incremental processing constraint effectively reduces the search space of the model.

This approach may be conceptually similar to our approach in that both try to reduce the search space of the model that comes from the constraint on human sentence processing (incremental left-to-right processing). However, his model, the grammar formalism (CCL), and the learning method are highly coupled with each other and it makes difficult to separate some component or idea in his framework for other applications. We instead investigate the effect of our structural constraint as a single component, which is much simpler and the idea can easily be portable to other applications.

Though CCL is similar to dependency, he evaluates the quality of the output on constituent-based bracketing scores by converting the CCL output to the equivalent constituent representation. We thus do not compare our approach to his method directly in this thesis.

CCG induction In the lexicalized grammar formalisms such as CCGs (Steedman, 2000), each nonterminal symbol in a parse tree encodes semantics about syntax and is not an arbitrary symbol unlike previous CFG-based grammar induction approaches. This observation motivates recent attempts for inducing CCGs with a small amount of supervision.

Bisk and Hockenmaier (2012) and Bisk and Hockenmaier (2013) present generative models over CCG trees and demonstrate that it achieves state-of-the-art scores on a number of languages in the WILS shared task dataset. For evaluation, after getting a CCG derivation, they extract dependencies by reading off predicate and argument (or modifier) structures encoded in CCG categories. Bisk and Hockenmaier (2015) present a model extension and thorough error analysis while Bisk et al. (2015) show how the idea can be applied when no identity on POS tags (e.g., whether a word cluster is VERB or not) is given with a small manual effort.

The key to the success of their approach is in the seed knowledge about category assignments for each input token. In CCGs or related formalisms, it is known that a parse tree is build almost deterministically if every category for input tokens are assigned correctly (Matsuzaki et al., 2007; Lewis and Steedman, 2014). In other words, the most difficult part in those parsers is the assignments of lexical categories, which highly restrict the ambiguity in the remaining parts. Bisk and Hockenmaier efficiently exploit this property of CCG by restricting possible CCG categories on POS tags. Their seed knowledge is that a sentence root should be a verb or a noun, and a noun should be an argument of a verb. They encode this knowledge to the model by seeding that only a verb can be assigned category S and only a noun can be assigned category N. Then, they expand the possible candidate categories for each POS tag in a bootstrap manner. For example, a POS tag next to a verb

| | | | |
|-------------|----------------|----------------|----------------|
| <i>The</i> | <i>man</i> | <i>ate</i> | <i>quickly</i> |
| DT | NNS | VBD | RB |
| N/N | N , S/S | S , N/N | S\S |
| (S/S)/(S/S) | (N\N)/(N\N) | S\N | (N\N)\(N\N) |
| | (N/N)\(N/N) | (S/S)\(S/S) | |
| | | (S\S)/(S\S) | |

Figure 2.27: An example of bootstrapping process for assigning category candidates in CCG induction borrowed from Bisk and Hockenmaier (2013). DT, NNS, VBD, and RB are POS tags. Bold categories are the initial seed knowledge, which is expanded by allowing the neighbor token to be a modifier.

is allowed to be assigned category $S\S$ or S/S , and so on.¹⁶ Figure 2.27 shows an example of this bootstrapping process, which we borrow from Bisk and Hockenmaier (2013).

After the process, the parameters of the generative model are learned using variational EM. During this phase, the category spanning the whole sentence is restricted to be S , or N if no verb exists in the sentence. This mechanism highly restricts the search space and allows efficient learning.

Finally, Garrette et al. (2015) explore another direction for learning CCG with small supervision. Unlike Bisk and Hockenmaier’s models that are based on gold POS tags, they try to learn the model from surface forms but with an incomplete tag dictionary mapping some words to possible categories. The essential difference between these two approaches is how to provide the seed knowledge to the model and it is an ongoing research topic (and probably one of the main goal in unsupervised grammar induction) to specify what kind of information should be given to the model and what can be learned from such seed knowledge.

2.4.5 Summary

This section surveyed the previous studies in unsupervised and lightly supervised grammar induction. As we have seen, dependency is the only structure that can be learned effectively with the well-studied techniques, e.g., PCFGs and the EM algorithm, except CCG, which may have a potential to replace this although the model tends to be inevitably more complex. For simplicity, our focus in thesis is dependency, but we argue that the success in dependency induction indicates that the idea could be extended to learning of the other grammars, e.g., CCG as well as more basic CFG-based constituent structures.

The key to the success of previous dependency-based approaches can be divided into the following categories:

Initialization The harmonic initializer is known to boost the performance and used in many previous models including Cohen and Smith (2009), Berg-Kirkpatrick et al. (2010), and Blunsom and Cohn (2010).

¹⁶The rewrite rules of CCG are defined by a small set of combinatory rules. For example, the rule $(S\backslash N)/N \rightarrow S\backslash N$ is an example of the forward application rule, which can be generally written as $X/Y \ Y \rightarrow X$. The backward application does the opposite: $Y \ X\backslash Y \rightarrow X$.

Principles of dependency The reducibility of Mareček and Žabokrtský (2012) and Mareček and Straka (2013) efficiently exploits the principle property in dependency and thus learning gets more stable.

Structural bias Smith and Eisner (2005) explores the effect of shorter dependency length bias, which is similar to the harmonic initialization but is more explicit.

Rules on POS tags Naseem et al. (2010) and Grave and Elhadad (2015) shows parameter-based constraints on POS tags can boost the performance. Søgaard (2012) is the evidence that such POS tag rules are already powerful in themselves to achieve reasonable scores.

The most relevant approach to ours that we present in Chapter 5 is the structural bias of Smith and Eisner (2005); However, as we have mentioned, they combine the technique with annealing and the selection of initialization method, which are tuned with the supervised model selection. Thus they do not explore the effect of a single structural bias, which is the main interest in our experiments. As another baseline, we also compare the performance with harmonic initialized models. The reducibility and rules on POS tags possibly have orthogonal effects to the structural bias. We will explore a small number of rules and see the combination effects with our structural constraints to get insights on the effect of our constraint when some amount of external supervision is provided.

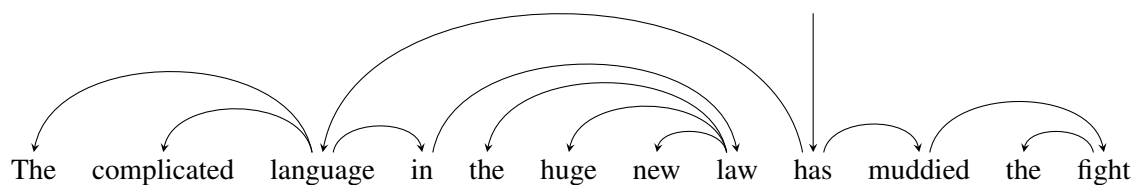
Chapter 3

Multilingual Dependency Corpora

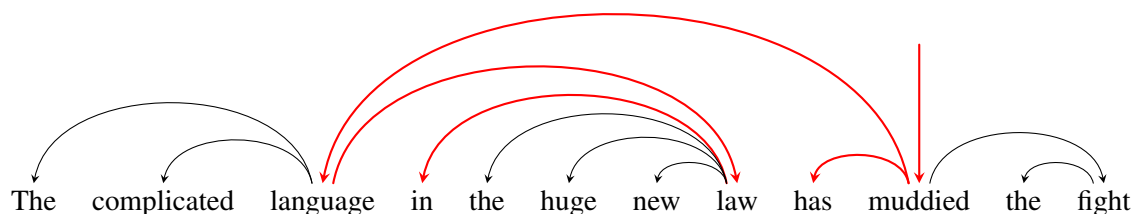
Cross-linguality is an important concept in this thesis. In the following chapters, we explore syntactic regularities or universals that exist in languages in several ways including corpus analyses, a supervised parsing study (Chapter 4), and an unsupervised parsing study (Chapter 5). All these studies were made possible by recent efforts for the development of multilingual corpora. This chapter summarizes the properties and statistics of the dataset we use in our experiments.

First, we survey the problem of the ambiguity in the definitions of *head* that we noted when introducing dependency grammars in Section 2.1.3. This problem is critical for our purpose; for example, if our unsupervised induction system performs so badly for a particular language, we do not know whether the reason is in the (possibly distinguished) annotation style or the inherent difficulty of that language (see also Section 5.3.3). In particular, we describe the duality of head, i.e., *function* head and *content* head, which is the main source of the reason why there can be several dependency representations for a particular syntactic construction.

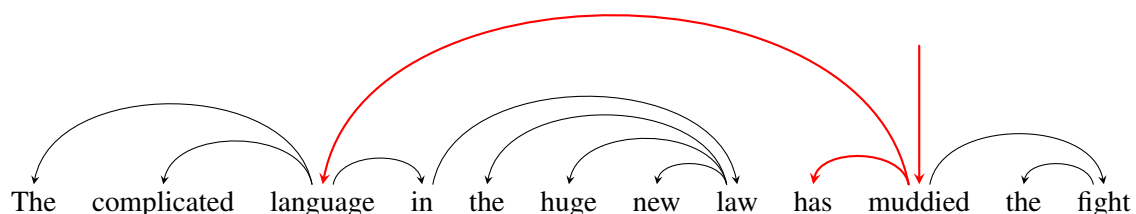
We then summarize the characteristics of the treebanks that we use. The first dataset, CoNLL shared tasks dataset (Buchholz and Marsi, 2006; Nivre et al., 2007a) is the first large collection of multilingual dependency treebanks (19 languages in total) in the literature, although is just a collection of existing treebanks and lacking annotation consistency across languages. This dataset thus may not fully adequate for our cross-linguistic studies. We will introduce this dataset and use it in our experiments mainly because it was our primary dataset in the preliminary version of the current study (Noji and Miyao, 2014), which was done when more adequate dataset such as Universal Dependencies (Marneffe et al., 2014) were not available. We use this dataset only for the experiments in Chapter 4. Universal Dependencies (UD) is a recent initiative to develop cross-linguistically consistent treebank annotation for many languages (Nivre, 2015). We choose this dataset as our primary resource for cross-linguistic experiments since currently it seems the best dataset that keeps the balance between the typological diversity in terms of the number of languages or language families and the annotation consistency. We finally introduce another recent annotation project called Google universal treebanks (McDonald et al., 2013). We use this dataset only for our unsupervised parsing experiments in Chapter 5 mainly for comparing the performance of our model with the current state-of-the-art systems. This dataset is a preliminary version of UD, so its data size and consistency is inferior. We summarize the major differences of approaches in two corpora in Section 3.4.



(a) Analysis on CoNLL dataset.



(b) Analysis on Stanford universal dependencies (UD).



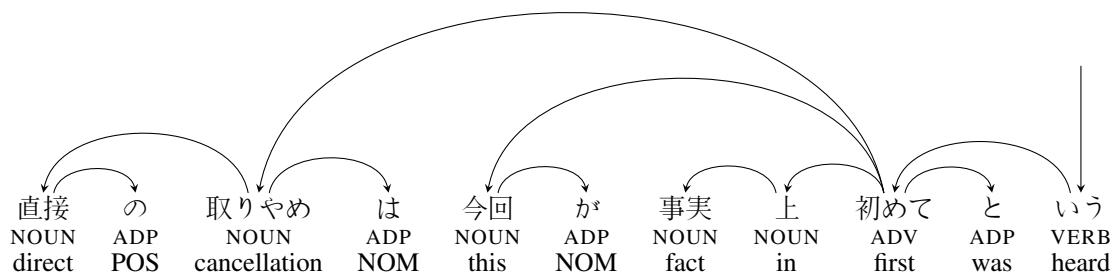
(c) Analysis on Google universal treebanks.

Figure 3.1: Each dataset that we use employs the different kind of annotation style. Bold arcs are ones that do not exist in the CoNLL style tree (a).

3.1 Heads in Dependency Grammars

Let us first see the examples. Figure 3.1 shows how the analysis of an English sentence would be changed across the datasets we use. Every analysis is *correct* under some linguistic theory. We can see that two analyses between the CoNLL style (Figure 3.1(a)) and the UD style (Figure 3.1(b)) are largely different, in particular around function words (e.g., *in* and *has*).

Function and content heads Zwicky (1993) argues that there is a duality in the notion of heads, namely, function heads and content heads. In the view of function heads, the head of each constituent is the word that determines the syntactic role of it. The CoNLL style tree is largely function head-based; For example, the head in constituent “in the huge new law” in Figure 3.1(a) is “in”, since this preposition determines the syntactic role of the phrase (i.e., prepositional phrase modifying another noun or verb phrase). The construction of “has muddied” is similar; In the syntactic



I heard that this was in fact the first time of the direct cancellation.

Figure 3.2: A dependency tree in the Japanese UD. NOUN, ADV, VERB, and ADP are assigned POS tags.

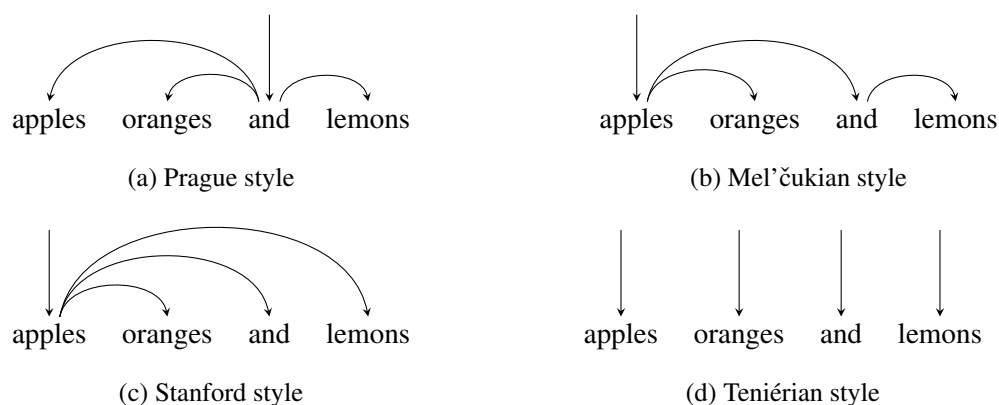


Figure 3.3: Four styles of annotation for coordination.

view, the auxiliary “has” becomes the head since it is this word that determines the aspect of this sentence (present perfect).

In another view of content heads, the head of each constituent is selected to be the word that most contributes to the semantics of it. This is the design followed in the UD scheme (Nivre, 2015). For example, in Figure 3.1(b) the head of constituent “in the huge new law” is the noun “law” instead of the preposition. Thus, in UD, every dependency arc is basically from a content word (head) to another content or function word (dependent). Figure 3.2 shows an example of sentence in Japanese treebank of UD. We can see that every function word (e.g., ADP) is attached to some content word, such as NOUN and ADV (adverb).

Other variations Another famous construction that has several variations in analysis is coordination, which is inherently *multiple-head* construction and is difficult to deal with in dependency. Popel et al. (2013) give detailed analysis of coordination structures employed in several existing treebanks. There are roughly four families of approaches (Zeman et al., 2014) in existing treebanks

as shown in Figure 3.3. Each annotation style has the following properties:

Prague All conjuncts are headed by the conjunction (Hajič et al., 2006).

Mel’čukian The first/last conjunct is the head, and others are organized in a chain (Mel’čuk, 1988).

Stanford The first conjunct is the head, others are attached directly to it (de Marneffe and Manning, 2008).

Teniérian There is no common head, and all conjuncts are attached directly to the node modified by the coordination structure (Tesnière, 1959).

Note that through our experiments we do not make any claims on which annotation style is the most appropriate for dependency analysis. In other words, we do not want to commit to a particular linguistic theory. The main reason why we focus on UD is that it is the dataset with the highest annotation consistency across languages now available, as we describe in the following.

3.2 CoNLL Shared Tasks Dataset

This dataset consists of 19 language treebanks used in the CoNLL shared tasks 2006 (Buchholz and Marsi, 2006) and 2007 (Nivre et al., 2007a), in which the task was the multilingual supervised dependency parsing. See the list of languages and statistics in Table 3.1. There is generally no annotation consistency across languages in various constructions. For example, the four types of coordination annotation styles all appear in this dataset; Prague style is used in, e.g., Arabic, Czech, and English, while Mel’čukian style is found in, e.g., German, Japanese, and Swedish, etc. Function and content head choices are also mixed across languages as well as the constructions in each language. For example, in English, the basic style is function head-based while some exceptions are found in e.g., infinitive marker in a verb phrase, such as “... allow executives to report ...” in which the head of “to” is “report” instead of “allow”. The idea of regarding a determiner as a head is the extreme of function head-based view (Abney, 1987; Hudson, 2004), and most treebanks treat a noun as a head while the determiner head is also employed in some treebank, such as Danish. Zeman et al. (2014) gives more detailed survey on the differences of annotation styles in this dataset.

The dataset consists of the following treebanks. Note that some languages (Arabic, Chinese, Czech, and Turkish) are used in both 2006 and 2007 shared tasks in different versions; in which case we use only 2007 data. Also a number of treebanks, such as Basque, Chinese, English, etc, are annotated originally in phrase-structure trees, which are converted to dependency trees with heuristics rules extracting a head token from each constituent.

Arabic: Prague Arabic Dependency Treebank 1.0 (Srnč et al., 2008).

Basque: 3LB Basque treebank (Aduriz et al., 2003).

Bulgarian: BulTreeBank (Simov and Osenova, 2005).

Catalan: The Catalan section of the CESS-ECE Syntactically and Semantically Annotated Corpora (M. Antónia Martí and Bertran, 2007).

| Language | #Sents. | #Tokens | | | | Punc. (%) | Av. len. |
|------------|---------|-----------|-----------|-----------|---------------|--------------|----------|
| | | ≤ 10 | ≤ 15 | ≤ 20 | $\leq \infty$ | | |
| Arabic | 3,043 | 2,833 | 5,193 | 8,656 | 116,793 | 8.3 | 38.3 |
| Basque | 3,523 | 7,865 | 19,351 | 31,384 | 55,874 | 18.5 | 15.8 |
| Bulgarian | 13,221 | 34,840 | 75,530 | 114,687 | 196,151 | 14.3 | 14.8 |
| Catalan | 15,125 | 9,943 | 31,020 | 66,487 | 435,860 | 11.6 | 28.8 |
| Chinese | 57,647 | 269,772 | 326,275 | 337,908 | 342,336 | 0.0 | 5.9 |
| Czech | 25,650 | 48,452 | 110,516 | 191,635 | 437,020 | 14.7 | 17.0 |
| Danish | 5,512 | 10,089 | 24,432 | 40,221 | 100,238 | 13.9 | 18.1 |
| Dutch | 13,735 | 40,816 | 75,665 | 110,118 | 200,654 | 11.2 | 14.6 |
| English | 18,791 | 13,969 | 47,711 | 106,085 | 451,576 | 12.2 | 24.0 |
| German | 39,573 | 66,741 | 164,738 | 292,769 | 705,304 | 13.5 | 17.8 |
| Greek | 2,902 | 2,851 | 8,160 | 16,076 | 70,223 | 10.1 | 24.1 |
| Hungarian | 6,424 | 8,896 | 23,676 | 42,796 | 139,143 | 15.5 | 21.6 |
| Italian | 3,359 | 5,035 | 12,350 | 21,599 | 76,295 | 14.7 | 22.7 |
| Japanese | 17,753 | 52,399 | 81,561 | 105,250 | 157,172 | 11.6 | 8.8 |
| Portuguese | 9,359 | 13,031 | 30,060 | 54,804 | 212,545 | 14.0 | 22.7 |
| Slovene | 1,936 | 4,322 | 9,647 | 15,555 | 35,140 | 18.0 | 18.1 |
| Spanish | 3,512 | 3,968 | 9,716 | 18,007 | 95,028 | 12.5 | 27.0 |
| Swedish | 11,431 | 20,946 | 55,670 | 96,907 | 197,123 | 10.9 | 17.2 |
| Turkish | 5,935 | 21,438 | 34,449 | 44,110 | 69,695 | 16.0 | 11.7 |

Table 3.1: Overview of CoNLL dataset (mix of training and test sets). Punc. is the ratio of punctuation tokens in a whole corpus. Av. len. is the average length of a sentence.

Chinese: Sinica treebank (Chen et al., 2003).

Czech: Prague Dependency Treebank 2.0 (Hajič et al., 2006).

Danish: Danish Dependency Treebank (Kromann et al., 2004).

Dutch: Alpino treebank (van der Beek et al., 2002).

English: The Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993).

German: TIGER treebank (Brants et al., 2002).

Greek: Greek Dependency Treebank (Prokopidis et al., 2005).

Hungarian: Szeged treebank (Csendes et al., 2005).

Italian: A subset of the balanced section of the Italian SyntacticSemantic Treebank (Montemagni et al., 2003).

Japanese: Japanese Verbmobil treebank (Kawata and Bartels, 2000). This is mainly the collection of speech conversations and thus the average length is relatively short.

Portuguese: The Bosque part of the Floresta sintá(c)tica (Afonso et al., 2002) covering both Brazilian and European Portuguese.

Slovene: Slovene Dependency Treebank (Džeroski et al., 2006).

Spanish: Cast3LB (Civit and Martí, 2004).

Swedish: Talbanken05 (Nivre et al., 2006).

Turkish: METU-Sabancı Turkish Treebank used in CoNLL 2007 (Atalay et al., 2003).

3.3 Universal Dependencies

UD is a collection of treebanks each of which is designed to follow the annotation guideline based on the Stanford typed dependencies (de Marneffe and Manning, 2008), which is in most cases content head-based as we mentioned in Section 3.1. We basically use the version 1.1 of this dataset, from which we exclude Finnish-FTB since UD also contains another Finnish treebank, which is larger, and add Japanese, which is included in version 1.2 dataset first. Typically a treebank is created by first transforming trees in an existing treebank with some script into the trees to follow the annotation guideline, and then manually correcting the errors.

Another characteristic of this dataset is the set of POS tags and dependency labels are consistent across languages. Appendix B summarizes the POS tagset of UD. We do not discuss dependency labels since we omit them.

Below is the list of sources of treebanks. We omit the languages if the source is the same as the CoNLL dataset described above. Note the source of some languages, such as English and Japanese, are changed from the previous dataset. See Table 3.2 for the list of all 19 languages as well as the statistics.

Croatian: SETimes.HR (Agić and Ljubešić, 2014).

English: English Web (Silveira et al., 2014).

Finnish: Turku Dependency Treebank (Haverinen et al., 2010).

German: Google universal treebanks (see Section 3.4).

Hebrew: Hebrew Dependency Treebank (Goldberg, 2011).

Indonesian: Google universal treebanks (see Section 3.4).

Irish: Irish Dependency Treebank (Lynn et al., 2014).

Japanese : Kyoto University Text Corpus 4.0 (Kawahara et al., 2002; Kanayama et al., 2015).

Persian : Uppsala Persian Dependency Treebank (Seraji, 2015).

Spanish : Google universal treebanks (see Section 3.4).

| Language | #Sents. | #Tokens | | | | Punc. | Av. len. |
|------------|---------|-----------|-----------|-----------|---------------|-------|----------|
| | | ≤ 10 | ≤ 15 | ≤ 20 | $\leq \infty$ | | |
| Basque | 5,273 | 19,597 | 38,612 | 51,305 | 60,563 | 17.3 | 11.4 |
| Bulgarian | 9,405 | 27,903 | 58,386 | 84,318 | 125,592 | 14.3 | 13.3 |
| Croatian | 3,957 | 3,850 | 12,718 | 26,614 | 87,765 | 12.9 | 22.1 |
| Czech | 87,913 | 160,930 | 377,994 | 654,559 | 1,506,490 | 14.6 | 17.1 |
| Danish | 5,512 | 10,089 | 24,432 | 40,221 | 100,238 | 13.8 | 18.1 |
| English | 16,622 | 36,189 | 74,361 | 115,511 | 254,830 | 11.7 | 15.3 |
| Finnish | 13,581 | 39,797 | 85,601 | 123,036 | 181,022 | 14.6 | 13.3 |
| French | 16,468 | 13,988 | 51,525 | 106,303 | 400,627 | 11.1 | 24.3 |
| German | 15,918 | 24,418 | 74,400 | 135,117 | 298,614 | 13.0 | 18.7 |
| Greek | 2,411 | 2,229 | 6,707 | 13,493 | 59,156 | 10.6 | 24.5 |
| Hebrew | 6,216 | 5,527 | 17,575 | 35,128 | 158,855 | 11.5 | 25.5 |
| Hungarian | 1,299 | 1,652 | 5,196 | 9,913 | 26,538 | 14.6 | 20.4 |
| Indonesian | 5,593 | 6,890 | 23,009 | 42,749 | 121,923 | 14.9 | 21.7 |
| Irish | 1,020 | 1,901 | 3,695 | 6,202 | 23,686 | 10.6 | 23.2 |
| Italian | 12,330 | 24,230 | 51,033 | 79,901 | 277,209 | 11.2 | 22.4 |
| Japanese | 9,995 | 6,832 | 24,657 | 54,395 | 267,631 | 10.8 | 26.7 |
| Persian | 6,000 | 6,808 | 18,011 | 34,191 | 152,918 | 8.7 | 25.4 |
| Spanish | 16,006 | 10,489 | 40,087 | 88,665 | 432,651 | 11.0 | 27.0 |
| Swedish | 6,026 | 13,045 | 31,343 | 51,333 | 96,819 | 10.7 | 16.0 |

Table 3.2: Overview of UD dataset (mix of train/dev/test sets). Punc. is the ratio of punctuation tokens in a whole corpus. Av. len. is the average length of a sentence.

3.4 Google Universal Treebanks

This dataset is a collection of 12 languages treebanks, i.e., Brazilian-Portuguese, English, Finnish, French, German, Italian, Indonesian, Japanese, Korean, Spanish and Swedish. Most treebanks are created by hand in this project except the following two languages:

English: Automatically convert from the Wall Street Journal portion of the Penn Treebank (Marcus et al., 1993) (with a different conversion method than the CoNLL dataset).

Swedish: Talbanken05 (Nivre et al., 2006) as in CoNLL dataset.

Basically every treebank follow the annotation guideline based on the Stanford typed dependencies as in UD, but contrary to UD, the annotation of Google treebanks is not fully content head-based. As we show in Figure 3.1(c), it annotates specific constructions in function head-based, in particular ADP phrases.

We do not summarize the statistics of this dataset here as we use it only in our experiments in Chapter 5 where we will see the statistics of the subset of the data that we use (see Section 5.3.1).

Chapter 4

Left-corner Transition-based Dependency Parsing

Based on several recipes introduced in Chapter 2, we now build a left-corner parsing algorithm operating on dependency grammars. In this chapter, we formalize the algorithm as a transition system for dependency parsing (Nivre, 2008) that roughly corresponds to the dependency version of a push-down automaton (PDA).

We have introduced PDAs with the left-corner parsing strategy for CFGs (Section 2.2.3) as well as a conversion method of any projective dependency trees into an equivalent CFG parse (Section 2.1.4). Thus one may suspect that it is straightforward to obtain a left-corner parsing algorithm for dependency grammars by, e.g., developing a CFG parser that will build a CFG parse encoding dependency information at each nonterminal symbol.

In this chapter, however, we take a different approach to build an algorithm in a non-trivial way. One reason for this is because such a CFG-based approach cannot be an *incremental* algorithm. On the other hand, our algorithm in this chapter is incremental; that is, it can construct a partial parse on the stack, without seeing the future input tokens. Incrementality is important for assessing parser performance with a comparison to other existing parsing methods, which basically assume incremental processing. We perform such empirical comparison in Sections 4.4 and 4.5.

For example, let us assume to build a parse in Figure 4.1(b), which corresponds to the CFG parse for a dependency tree on “a big dog”. To recognize this parse on the left-corner PDA in Section 2.2.3, after shifting token “a” (which becomes $X[a]$), the PDA may covert it to the symbol “ $X[\text{dog}]/X[\text{dog}]$ ”. However, for creating such a symbol, we have to know that “dog” will appear on the remaining inputs at this point, which is impossible in incremental parsing. This contrasts with the left-corner parser for phrase-structure grammars that we considered in Section 2.2.3 in which there is only a finite inventory of nonterminal, which might be predicted.

The algorithm we formalize in this chapter does not introduce such symbols to enable incremental parsing. We do so by introducing a new concept, a *dummy* node, which efficiently abstracts the predicted structure of a subtree in a compact way. Another important point is since this algorithm directly operates on a dependency tree (not via a CFG form), we can get intuition into how the left-corner parser builds a dependency parse tree. This becomes important when developing efficient

tabulating algorithm with head-splitting (Section 2.3.5) in Chapter 5.

We formally define our algorithm as a *transition system*, a stack-based formalization like push-down automata and is the most popular way for obtaining algorithms for dependency grammars (Nivre, 2003; Yamada and Matsumoto, 2003; Nivre, 2008; Gómez-Rodríguez and Nivre, 2013). As we discussed in Section 2.2.3, a left-corner parser can capture the degree of center-embedding of a construction by its stack depth. Our algorithm preserves this property, and its stack depth increases only when processing dependency structures involving center-embedding.

The empirical part of this chapter comprises of two kinds of experiments: First, we perform a corpus analysis to show that our left-corner algorithm consistently requires less stack depth to recognize annotated trees relative to other algorithms across languages. The result also suggests the existence of a syntactic universal by which deeper center-embedding is a rare construction across languages, which has not yet been quantitatively examined cross-linguistically. The second experiment is a supervised parsing experiment, which can be seen as an alternative way to assess the parser’s ability to capture important syntactic regularities. In particular, we will find that the parser using our left-corner algorithm is consistently less sensitive to the decoding constraints of stack depth bound across languages. Conversely, the performance of other dependency parsers such as the arc-eager parser is largely affected by the same constraints.

The motivation behind these comparisons is to examine whether the stack depth of a left-corner parser is in fact a meaningful measure to explain the syntactic universal among other alternatives, which would be valuable for other applications such as unsupervised grammar induction that we explore in Chapter 5.

The first experiment is a *static* analysis, which strictly analyzes the observed tree forms in the treebanks, while the second experiment takes *parsing errors* into account. Though the result of the first experiment seems clearer to claim a universal property of language, the result of the second experiment might also be important for real applications. Specifically we will find that the rate of performance drop with a decoding constraint is smaller than the expected value from the coverage result of the first experiment. This suggests that a good approximation of the observed syntactic structures in treebanks is available from a highly restricted space if we allow small portion of parse errors. Since real applications always suffer from parse errors, this result is more appealing for finding a good constraint to restrict the possible tree structures.

This chapter proceeds as follows: Since our empirical concern is the relative performance of our left-corner algorithm compared to existing transition-based algorithms, we begin the discussion in this chapter with a survey of stack depth behavior in existing algorithms in Section 4.2. This discussion is an extension of a preliminary survey about the incrementality of transition systems by Nivre (2004), which is (to our knowledge) the only study discussing how stack elements increase for a particular dependency structures in some algorithm. Then, in Section 4.3, we develop our new transition system that follows a left-corner parsing strategy for dependency grammars and discuss the formal properties of the system, such as the spurious ambiguity of the system and its implications, which are closely relevant to the spurious ambiguity problem we discussed in Section 2.1.4. The empirical part is devoted to Sections 4.4 and 4.5, focusing on the static corpus analysis and supervised parsing experiments, respectively. Finally, we give discussion along with the the relevant previous studies in Section 4.6 to conclude this chapter.

The preliminary version of this chapter appeared as Noji and Miyao (2015), which was itself an

extension of Noji and Miyao (2014). Although these previous versions limited the dataset to the one in the CoNLL shared tasks (Buchholz and Marsi, 2006; Nivre et al., 2007a), we add new analysis on Universal dependencies (Marneffe et al., 2014) (see also Chapter 3). The total number of analyzed treebanks is 38 in total across 26 languages.

4.1 Notations

We first introduce several important concepts and notations used in this chapter.

Transition system Every parsing algorithm presented in this chapter can be formally defined as a transition system. The description below is rather informal; See Nivre (2008) for more details. A transition system is an abstract state machine that processes sentences and produces parse trees. It has a set of *configurations* and a set of *transition actions* applied to a configuration. Each system defines an *initial configuration* given an input sentence. The parsing process proceeds by repeatedly applying an action to the current configuration. After a finite number of transitions the system arrives at a *terminal configuration*, and the dependency tree is read off the terminal configuration.

Formally, each configuration is a tuple (σ, β, A) ; here, σ is a stack, and we use a vertical bar to signify the append operation, e.g., $\sigma = \sigma' | \sigma_1$ denotes σ_1 is the topmost element of stack σ . Further, β is an input buffer consisting of token indexes that have yet to be processed; here, $\beta = j | \beta'$ indicates that j is the first element of β . Finally, $A \subseteq V_w \times V_w$ is a set of arcs given V_w , a set of token indexes for sentence w .

Transition-based parser We distinguish two similar terms, a transition system and a transition-based parser in this chapter. A transition system formally characterizes how a tree is constructed via transitions between configurations. On the other hand, a parser is built on a transition system, and it selects the best *action sequence* (i.e., the best parse tree) for an input sentence probably with some scoring model. Since a transition system abstracts the way of constructing a parse tree, when we mention a *parsing algorithm*, it often refers to a transition system, not a parser. Most of the remaining parts of this chapter is about transition systems, except Section 4.5, in which we compare the performance of several parsers via supervised parsing experiments.

Center-embedded dependency structure The concept of center-embedding introduced in Section 2.2.1 is originally defined on a constituent structure, or a CFG parse. Remember that a dependency tree also encodes constituent structures implicitly (see Figure 4.1) but the conversion from a dependency tree into a CFG parse (in CNF) is not unique, i.e., there is a spurious ambiguity (see Section 2.1.4). This ambiguity implies there is a subtlety for defining the degree of center-embedding for a dependency structure.

We argue that the tree structure of a given dependency tree (i.e., whether it belongs to center-embedding) cannot be determined by a given tree itself; We can determine the tree structure of a dependency tree *only if* we have some one-to-one conversion method from a dependency tree to a CFG parse. For example some conversion method may always convert a tree of Figure 4.1(c) into

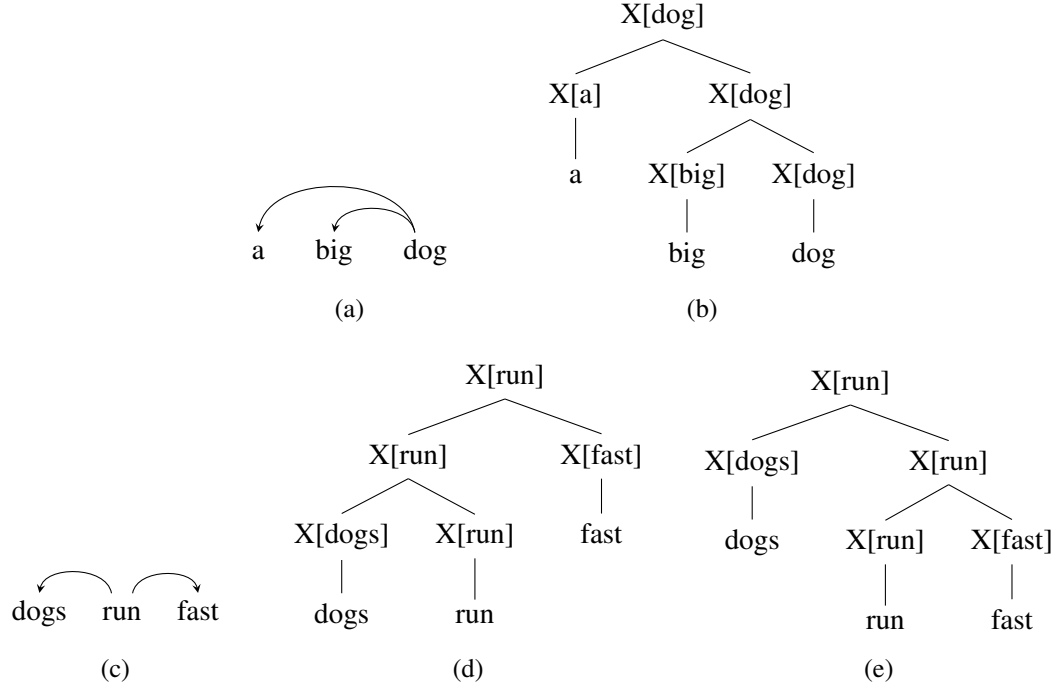


Figure 4.1: Conversions from dependency trees into CFG parses; (a) can be uniquely converted to (b), while (c) can be converted to both (d) and (e).

the one of Figure 4.1(d). In other words, the tree structure of a dependency tree should be discussed along with such a conversion method. We discuss this subtlety more in Section 4.3.3.

We avoid this ambiguity for a while by restricting our attention to the tree structures like Figure 4.1(a) in which we can obtain the corresponding CFG parse uniquely. For example the dependency tree in Figure 4.1(a) is an example of a *right-branching* dependency tree. Similarly we call a given dependency tree is center-embedding, or left- (right-)branching, depending on the implicit CFG parse when there is no conversion ambiguity.

4.2 Stack Depth of Existing Transition Systems

This section surveys how the stack depth of existing transition systems grows given a variety of dependency structures. These are used as baseline systems in our experiments in Sections 4.4 and 4.5.

4.2.1 Arc-standard

The arc-standard system (Nivre, 2004) consists of the following three transition actions, with (h, d) representing a dependency arc from h (head) to d (dependent).

- SHIFT: $(\sigma, j | \beta, A) \mapsto (\sigma | j, \beta, A)$;

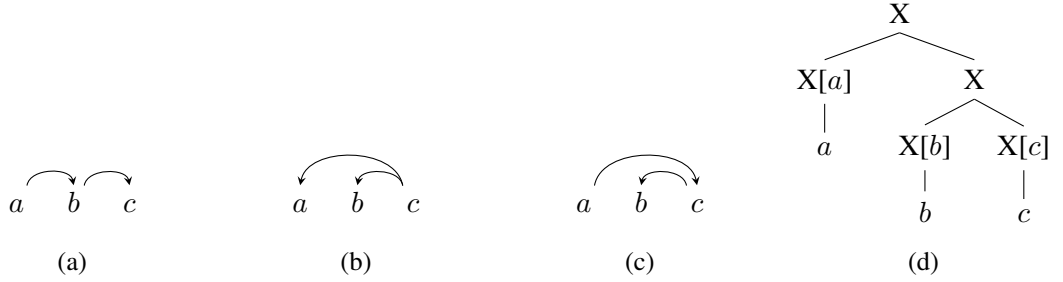


Figure 4.2: (a)-(c) Right-branching dependency trees for three words and (d) the corresponding CFG parse.

- LEFTARC: $(\sigma|\sigma'_2|\sigma'_1, \beta, A) \mapsto (\sigma|\sigma'_1, \beta, A \cup \{(\sigma'_1, \sigma'_2)\})$;
- RIGHTARC: $(\sigma|\sigma'_2|\sigma'_1, \beta, A) \mapsto (\sigma|\sigma'_2, \beta, A \cup \{(\sigma'_2, \sigma'_1)\})$.

We first observe here that the stack depth of the arc-standard system increases linearly for a right-branching structure, such as $a \frown b \frown c \dots$, in which the system first shifts all words on the stack before connecting each pair of words. Nivre (2004) analyzed this system and observed that stack depth grows when processing a dependency tree that becomes right-branching with a CFG conversion. Figure 4.2 shows these dependency trees for three words; the system must construct a subtree of b and c before connecting a to either, thus increasing stack depth. This occurs because the system builds a tree in a bottom-up manner, i.e., each token collects all dependents before being attached to its head. The arc-standard system is essentially equivalent to the push-down automaton of a CFG in CNF with a bottom-up strategy (Nivre, 2004), so it has the same property as the bottom-up parser for a CFG. This equivalence also indicates that its stack depth increases for center-embedded structures.

4.2.2 Arc-eager

The arc-eager system (Nivre, 2003) uses the following four transition actions:

- SHIFT: $(\sigma, j|\beta, A) \mapsto (\sigma|j, \beta, A)$;
- LEFTARC: $(\sigma|\sigma'_1, j|\beta, A) \mapsto (\sigma, j|\beta, A \cup \{(j, \sigma'_1)\})$ (if $\neg \exists k, (k, \sigma'_1) \in A$);
- RIGHTARC: $(\sigma|\sigma'_1, j|\beta, A) \mapsto (\sigma|\sigma'_1|j, \beta, A \cup \{(\sigma'_1, j)\})$;
- REDUCE: $(\sigma|\sigma'_1, \beta, A) \mapsto (\sigma, \beta, A)$ (if $\exists k, (k, \sigma'_1) \in A$).

Note that LEFTARC and REDUCE are not always applicable. LEFTARC requires that σ'_1 is not a dependent of any other tokens, while REDUCE requires that σ'_1 is a dependent of some token (attached to its head). These conditions originate from the property of the arc-eager system by which each element on the stack may not be disjoint. In this system, two successive tokens on the stack may be combined with a left-to-right arc, i.e., $a \frown b$, thus constituting a *connected component*.

| | Left-branching | Right-branching | Center-embedding |
|--------------|----------------|-----------------|------------------|
| Arc-standard | $O(1)$ | $O(n)$ | $O(n)$ |
| Arc-eager | $O(1)$ | $O(1 \sim n)$ | $O(1 \sim n)$ |
| Left-corner | $O(1)$ | $O(1)$ | $O(n)$ |

Table 4.1: Order of required stack depth for each structure for each transition system. $O(1 \sim n)$ means that it recognizes a subset of structures within a constant stack depth but demands linear stack depth for the other structures.

For this system, we slightly abuse the notation and define stack depth as the number of connected components, not as the number of tokens on the stack, since our concern is the syntactic bias that may be captured with measures on the stack. With the definition based on the number of tokens on the stack, the arc-eager system would have the same stack depth properties as the arc-standard system. As we see below, the arc-eager approach has several interesting properties with this modified definition.¹

From this definition, unlike the arc-standard system, the arc-eager system recognizes the structure shown in Figure 4.2(a) and more generally $a \frown b \frown c \frown \dots$ within constant depth (just one) since it can connect all tokens on the stack with consecutive RIGHTARC actions. More generally, the stack depth of the arc-eager system never increases as long as all dependency arcs are left to right. This result indicates that the construction of the arc-eager system is no longer purely bottom-up and makes it difficult to formally characterize the stack depth properties based on the tree structure.

We argue two points regarding the stack depth of the arc-eager system. First, it recognizes a subset of the right-branching structures within a constant depth, as we analyzed above, while increasing stack depth linearly for other right-branching structures, including the trees shown in Figures 4.2(b) and 4.2(c). Second, it recognizes a subset of the center-embedded structures within a constant depth, such as $a \frown b \frown c \frown d$, which becomes center-embedded when converted to a constituent tree with all arcs left-to-right. For other center-embedded structures, the stack depth grows linearly as with the arc-standard system.

We summarize the above results in Table 4.1. The left-corner transition system that we propose next has the properties of the third row of the table, and its stack depth grows only on center-embedded dependency structures.

4.2.3 Other systems

All systems in which stack elements cannot be connected have the same properties as the arc-standard system because of their bottom-up constructions including the hybrid system of Kuhlmann et al. (2011). Kitagawa and Tanaka-Ishii (2010) and Sartorio et al. (2013) present an interesting variant that attaches one node to another node that may not be the head of a subtree on the stack. We do not explore these systems in our experiments because their stack depth essentially has the same properties as the arc-eager system, e.g., their stack depth does not always grow on center-embedded structures, although it grows on some kinds of right-branching structures.

¹ The stack of the arc-eager system can be seen as the stack of stacks; i.e., each stack element itself is a stack preserving a connected subtree (a right spine). Our definition of stack depth corresponds to the depth of this stack of stacks.

4.3 Left-corner Dependency Parsing

In this section, we develop our dependency transition system with the left-corner strategy. Our starting point is the push-down automaton for a CFG that we developed in Section 2.2.3. We will describe how the idea in this automaton can be extended for dependency trees by introducing the concept of *dummy nodes* that abstract the prediction mechanism required to achieve the left-corner parsing strategy.

4.3.1 Dummy node

The key characteristic of our transition system is the introduction of a dummy node in a subtree, which is needed to represent a subtree containing predicted structures, such as the symbol A/B in Figure 2.12, which predicts an existence of B top-down. To intuitively understand the parser actions, we present a simulation of transitions for the sentence shown in Figure 4.2(b) for which all existing systems demand a linear stack depth. Our system first shifts a and then conducts a *prediction* operation that yields subtree $a \leftarrow x$, where x is a dummy node. Here, we predict that a will become a left dependent of an incoming word. Next, it shifts b to the stack and then conducts a *composition* operation to obtain a tree $a \leftarrow b \leftarrow x$. Finally, c is inserted into the position of x , thus recovering the tree.

4.3.2 Transition system

Our system uses the same notation for a configuration as other systems presented in Section 4.2. Figure 4.3 shows an example of a configuration in which the i -th word in a sentence is written as w_i on the stack. Each element on the stack is a list representing a right spine of a subtree, which is similar to Kitagawa and Tanaka-Ishii (2010) and Sartorio et al. (2013). Here, right spine $\sigma_i = \langle \sigma_{i1}, \sigma_{i2}, \dots, \sigma_{ik} \rangle$ consists of all nodes in a descending path from the head of σ_i , i.e., from σ_{i1} , taking the rightmost child at each step. We also write $\sigma_i = \sigma'_i | \sigma_{ik}$, meaning that σ_{ik} is the rightmost node of spine σ_i . Each element of σ_i is an index of a token in a sentence or a subtree rooted at a dummy node, $x(\lambda)$, where λ is the set of left dependents of x . We state that right spine σ_i is *complete* if it does not contain any dummy nodes, while σ_i is *incomplete* if it contains a dummy node.²

All transition actions in our system are defined in Figure 4.4. INSERT is essentially the same as the SCAN operation in the original left-corner PDA for CFGs (Figure 2.12). Other changes are that we divide PREDICTION and COMPOSITION into two actions, left and right. As in the left-corner PDA, by a shift action, we mean SHIFT or INSERT, while a reduce action means one of prediction and composition actions.

² σ_i with a dummy node corresponds to a stack symbol of the form A/B in the left-corner PDA, which we called incomplete in Section 2.2.3. Thus, the meaning of these notions (i.e., complete and incomplete) is the same in two algorithms. The main reason for us to use spine-based notation stems from our use of a dummy node, which postpones the realization of dependency arcs connected to it. To add arcs appropriately to A when a dummy is filled with a token, it is necessary to keep the surrounding information of the dummy node (this occurs in INSERT and RIGHTCOMP), which can be naturally traced by remembering each right spine.

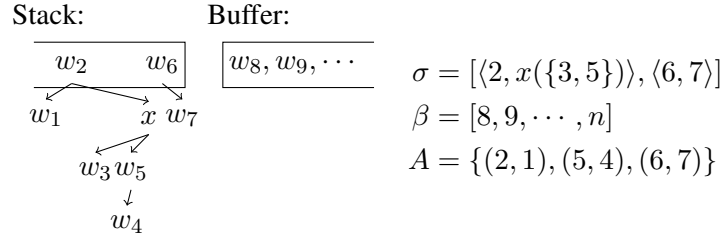


Figure 4.3: Example configuration of a left-corner transition system.

| | |
|-----------|---|
| SHIFT | $(\sigma, j \beta, A) \mapsto (\sigma \langle j \rangle, \beta, A)$ |
| INSERT | $(\sigma \langle \sigma'_1 i x(\lambda) \rangle), j \beta, A \mapsto (\sigma \langle \sigma'_1 i j \rangle, \beta, A \cup \{(i, j)\} \cup \{\cup_{k \in \lambda} (j, k)\})$ |
| LEFTPRED | $(\sigma \langle \sigma_{11}, \dots \rangle, \beta, A) \mapsto (\sigma \langle x(\sigma_{11}) \rangle, \beta, A)$ |
| RIGHTPRED | $(\sigma \langle \sigma_{11}, \dots \rangle, \beta, A) \mapsto (\sigma \langle \sigma_{11}, x(\emptyset) \rangle, \beta, A)$ |
| LEFTCOMP | $(\sigma \langle \sigma'_2 x(\lambda) \rangle \langle \sigma_{11}, \dots \rangle, \beta, A) \mapsto (\sigma \langle \sigma'_2 x(\lambda \cup \{\sigma_{11}\}) \rangle, \beta, A)$ |
| RIGHTCOMP | $(\sigma \langle \sigma'_2 x(\lambda) \rangle \langle \sigma_{11}, \dots \rangle, \beta, A) \mapsto (\sigma \langle \sigma'_2 \sigma_{11} x(\emptyset) \rangle, \beta, A \cup \{\cup_{k \in \lambda} (\sigma_{11}, k)\})$ |

Figure 4.4: Actions of the left-corner transition system including two shift operations (top) and reduce operations (bottom).

Shift Action As in the left-corner PDA, SHIFT moves a token from the top of the buffer to the stack. INSERT corresponds to the SCAN operation of the PDA, and replaces a dummy node on the top of the stack with a token from the top of the buffer. Note that before doing a shift action, a dummy x can be replaced by any words, meaning that arcs from and to x are unspecified. This is the key to achieve incremental parsing (see the beginning of this chapter). It is INSERT that these arcs are first specified, by filling the dummy node with an actual token. As in the left-corner PDA, the top element of the stack must be complete after a shift action.

Reduce Action As in the left-corner PDA, a reduce action is applied when the top element of the stack is complete, and changes it to an incomplete element.

LEFTPRED and RIGHTPRED correspond to PREDICTION in the left-corner PDA. Figure 4.5 describes the transparent relationships between them. LEFTPRED makes the head of the top stack element (i.e., σ_{11}) as a left dependent of a new dummy x , while RIGHTPRED predicts a dummy x as a right dependent of a . In these actions, if we think the original and resulting dependency forms in CFG, the correspondence to PREDICTION in the PDA is apparent. Specifically, the CFG forms of the resulting trees in both actions are the same. The only difference is the head label of the parent symbol, which is x in LEFTPRED while a in RIGHTPRED.

A similar correspondence holds between RIGHTCOMP, LEFTCOMP, and COMPOSITION in the PDA. We can interpret LEFTCOMP as two step operations as in COMPOSITION in the PDA (see Section 2.2.3): It first applies LEFTPRED to the top stack element, resulting in $a \leftarrow x$, and then unifies two x s to comprise a subtree. The connection to COMPOSITION is apparent from the figure. RIGHTCOMP, on the other hand, first applies RIGHTPRED to a , and then combines the resulting tree and the second top element on the stack. This step is a bit involved, which might be easier to

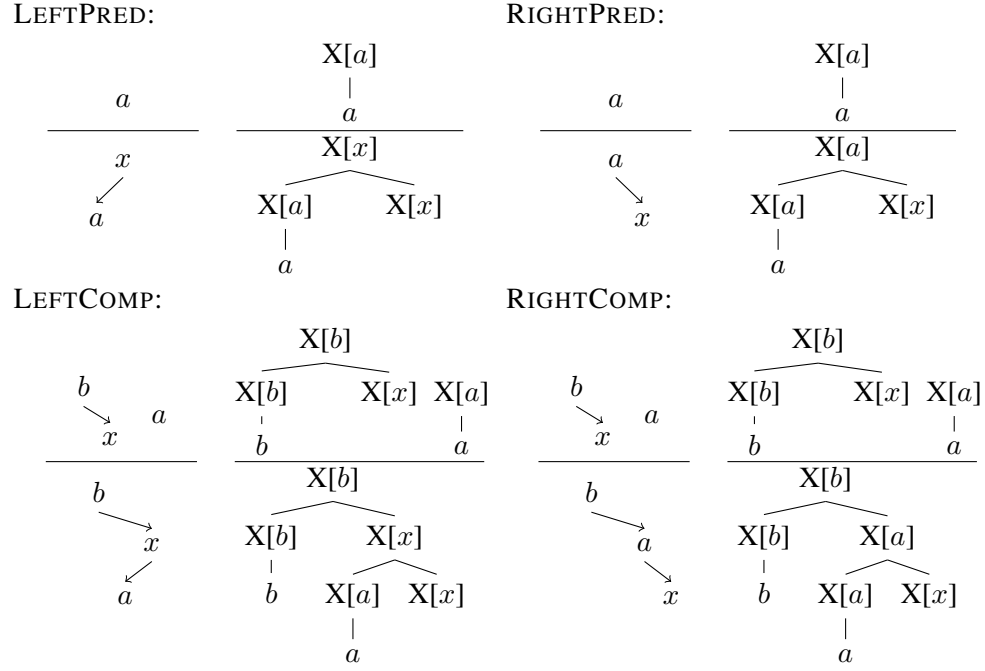


Figure 4.5: Correspondences of reduce actions between dependency and CFG. We only show minimal example subtrees for simplicity. However, a can have an arbitrary number of children, so can b or x , provided x is on a right spine and has no right children.

understand with the CFG form. On the CFG form, two unified nodes are $X[x]$, which is predicted top-down, and $X[a]$, which is recognized bottom-up (with the first RIGHTPRED step).³ Since x can be unified with any tokens, at this point, x in $X[x]$ is filled with a . Returning to dependency, this means that we insert the subtree rooted at a (after being applied RIGHTPRED) into the position of x in the second top element.

Note that from Figure 4.5, we can see that the dummy node x can only appear in a right spine of a CFG subtree. Now, we can reinterpret INSERT action on the CFG subtree, which attaches a token to a (predicted) preterminal $X[x]$, as in SCAN of the PDA, and then fills every x in a right spine with a shifted token. This can be seen as a kind of unification operation.

Relationship to the left-corner PDA As we have seen so far, though our transition system directly operates dependency trees, we can associate every step with a process to expand (partial) CFG parses as in the manner that the left-corner PDA would do.⁴ In every step, the transparency

³If the parent of $X[x]$ is also $X[x]$, all of them are filled with a recursively. This situation corresponds to the case in which dummy node x in the dependency tree has multiple left dependents, as in the resulting tree by LEFTCOMP in Figure 4.5.

⁴To be precise, we note that this CFG-based expansion process cannot be written in the form used in the left-corner PDA. For example, if we write items in LEFTCOMP and RIGHTCOMP in Figure 4.5 in the form of A/B , both results in the same transition: $X[b]/X[x] X[a] \mapsto X[b]/X[x]$. This is due to our use of a dummy node x , which plays different roles

| Step | Action | Stack (σ) | Buffer (β) | Set of arcs (A) |
|------|-----------|---|--------------------|--------------------------|
| | | ε | $a\ b\ c\ d$ | \emptyset |
| 1 | SHIFT | $\langle a \rangle$ | $b\ c\ d$ | \emptyset |
| 2 | RIGHTPRED | $\langle a, x(\emptyset) \rangle$ | $b\ c\ d$ | \emptyset |
| 3 | SHIFT | $\langle a, x(\emptyset) \rangle \langle b \rangle$ | $c\ d$ | \emptyset |
| 4 | RIGHTPRED | $\langle a, x(\emptyset) \rangle \langle b, x(\emptyset) \rangle$ | $c\ d$ | \emptyset |
| 5 | INSERT | $\langle a, x(\emptyset) \rangle \langle b, c \rangle$ | d | (b, c) |
| 6 | RIGHTCOMP | $\langle a, b, x(\emptyset) \rangle$ | d | $(b, c), (a, b)$ |
| 7 | INSERT | $\langle a, b, d \rangle$ | | $(b, c), (a, b), (b, d)$ |

Figure 4.6: An example parsing process of the left-corner transition system.

of two tree forms, i.e., dependency and CFG, is always preserved. This means that at the final configuration the CFG parse would be the one corresponding to the resulting dependency tree, and also at each step the stack depth is identical to the one that is incurred during parsing the final CFG parse with the original left-corner PDA. We will see this transparency with an example next. The connection between the stack depth and the degree of center-embedding, that we established in Theorem 2.1 for the PDA, also directly holds in this transition system. We restate this for our transition system in Section 4.3.4.

Example For an example, Figure 4.6 shows the transition of configurations during parsing a tree $a \frown b \frown c \frown d$, which corresponds to the parse in Figure 2.10(b) and thus involves one degree of center-embedding. Comparing to Figure 2.14, we can see that two transition sequences for the PDA (for CFG) and the transition system (for dependency) are essentially the same: the differences are that PREDICTION and COMPOSITION are changed to the corresponding actions (in this case, RIGHTPRED and RIGHCOMP) and SCAN is replaced with INSERT. This is essentially due to the transparent relationships between them that we discussed above. As in Figure 2.14, the stack depth two after a reduce action indicates center-embedding, which is step 4.

Other properties As in the left-corner PDA, this transition system also performs shift and reduce actions alternately (the proof is almost identical to the case of PDA). Also, given a sentence of length n , the number of actions required to arrive at the final configuration is $2n - 1$, because every token except the last word must be shifted once and reduced once, and the last word is always inserted as the final step.

Every projective dependency tree is derived from at least one transition sequence with this system, i.e., our system is *complete* for the class of projective dependency trees (Nivre, 2008). Though we omit the proof, this can be shown by appealing to the transparency between the transition system and the left-corner PDA, which is complete for a given CFG.

in two actions (e.g., RIGHCOMP assumes the first x is a) but the difference is lost with this notation. We thus claim that a CFG-based expansion step corresponds to a step in the left-corner PDA in that every action in the former expands the tree in the same way as the corresponding action of the left-corner PDA, as explained by Figure 4.5 (for reduce actions) and the body (for INSERT); the equivalence of SHIFT is apparent.

However, our system is *unsound* for the class of projective dependency trees, meaning that a transition sequence on a sentence does not always generate a valid projective dependency tree. We can easily verify this claim with an example. Let $a\ b\ c$ be a sentence and consider the action sequence “SHIFT LEFTPRED SHIFT LEFTPRED INSERT” with which we obtain the terminal configuration of $\sigma = [x(a), c]; \beta = []; A = \{(b, c)\}$ ⁵, but this is not a valid tree. The arc-eager system also suffers from a similar restriction (Nivre and Fernández-González, 2014), which may lead to lower parse accuracy. Instead of fixing this problem, in our parsing experiment, which is described in Section 4.5, we implement simple post-processing heuristics to combine those fragmented trees that remain on the stack.

4.3.3 Oracle and spurious ambiguity

This section presents and analyzes an oracle function for the transition system defined above. An oracle for a transition system is a function that returns a correct action given the current configuration and a set of gold arcs. The reasons why we develop and analyze the oracle are mainly two folds: First, we use this in our empirical corpus study in Section 4.4; that is, we analyze how stack depth increases during simulation of recovering dependency trees in the treebanks. Such simulation requires the method to extract the correct sequence of actions to recover the given tree. Second, we use it to obtain training examples for our supervised parsing experiments in Section 4.5. This is more typical reason to design the oracles for transition-based parsers (Nivre, 2008; Goldberg and Nivre, 2013).

Below we also point out the deep connection between the design of an oracle and a conversion process of a dependency into a (binary) CFG parse, which becomes the basis of the discussion in Section 4.3.4 on the degree of center-embedding of a given dependency tree, the problem we left in Section 4.1.

Since our system performs shift and reduce actions interchangeably, we need two functions to define the oracle. Let A_g be a set of arcs in the gold tree and c be the current configuration. We select the next shift action if the stack is empty (i.e., the initial configuration) or the top element of the stack is incomplete as follows:

- INSERT: Let $c = (\sigma | \langle \sigma'_1 | i | x(\lambda) \rangle, j | \beta, A)$. i may not exist. The condition is:
 - if i exists, $(i, j) \in A_g$ and j has no dependents in β ;
 - otherwise, $\exists k \in \lambda; (j, k) \in A_g$.
- SHIFT: otherwise.

If the top element on the stack is complete, we select the next reduce action as follows:

- LEFTCOMP: Let $c = (\sigma | \langle \sigma'_2 | i | x(\lambda) \rangle | \langle \sigma_{11}, \dots \rangle, \beta, A)$. i may not exist. Then
 - if i exists, σ_{11} has no dependents in β and i 's next dependent is the head of σ_{11} ;
 - otherwise, σ_{11} has no dependents in β and $k \in \lambda$ and σ_{11} share the same head.

⁵For clarity, we use words instead of indices for stack elements.

- **RIGHTCOMP**: Let $c = (\sigma | \langle \sigma'_2 | i | x(\lambda) \rangle | \langle \sigma_{11}, \dots \rangle, \beta, A)$. i may not exist. Then
 - if i exists, the rightmost dependent of σ_{11} is in β and $(i, \sigma_{11}) \in A_g$;
 - otherwise, the rightmost dependent of σ_{11} is in β and $\exists k \in \lambda, (\sigma_{11}, k) \in A_g$.
- **RIGHTPRED**: if $c = (\sigma | \langle \sigma_{11}, \dots \rangle, \beta, A)$ and σ_{11} has a dependent in β .
- **LEFTPRED**: otherwise.

Essentially, each condition ensures that we do not miss any gold arcs by performing the transition. This is ensured at each step so we can recover the gold tree in the terminal configuration. We use this oracle in our experiments in Sections 4.4 and 4.5.

Spurious ambiguity Next, we observe that the developed oracle above is not a unique function to return the gold action. Consider sentence $a^\wedge b^\wedge c$, which is a simplification of the sentence shown in Figure 4.1(c). If we apply the oracle presented above to this sentence, we obtain the following sequence:

$$\text{SHIFT} \rightarrow \text{LEFTPRED} \rightarrow \text{INSERT} \rightarrow \text{RIGHTPRED} \rightarrow \text{INSERT} \quad (4.1)$$

Note, however, that the following transitions also recover the parse tree:

$$\text{SHIFT} \rightarrow \text{LEFTPRED} \rightarrow \text{SHIFT} \rightarrow \text{RIGHTCOMP} \rightarrow \text{INSERT} \quad (4.2)$$

This is a kind of spurious ambiguities that we mentioned several times in this thesis (Sections 2.1.4, 2.3.5, and 4.1). Although in the transition-based parsing literature some works exist to improve parser performances by utilizing this ambiguity (Goldberg and Nivre, 2013) or by eliminating it (Hayashi et al., 2013), here we do not discuss such practical problems and instead *analyze* the differences in the transitions leading to the same tree.

Here we show that the spurious ambiguity of the transition system introduced above is essentially due to the spurious ambiguity of transforming a dependency tree into a CFG parse (Section 2.1.4). We can see this by comparing the implicitly recognized CFG parses with the two action sequences above. In sequence (4.1), **RIGHTPRED** is performed at step four, meaning that the recognized CFG parse has the form $((a \ b) \ c)$, while that of sequence (4.2) is $(a \ (b \ c))$ due to its **RIGHTCOMP** operation. This result indicates an oracle for our left-corner transition system implicitly binarizes a given gold dependency tree. The particular binarization mechanism associated with the oracle presented above is discussed next.

Implicit binarization We first note the property of the presented oracle that it follows the strategy of performing composition or insert operations when possible. As we saw in the given example, sometimes **INSERT** and **SHIFT** can both be valid for recovering the gold arcs, though here we always select **INSERT**. Sometimes the same ambiguity exists between **LEFTCOMP** and **LEFTPRED** or **RIGHTCOMP** and **RIGHTPRED**; we always prefer composition.

Then, we can show the following theorem about the binarization mechanism of this oracle.

Theorem 4.1. *The presented oracle above implicitly binarizes a dependency tree in the following manner:*

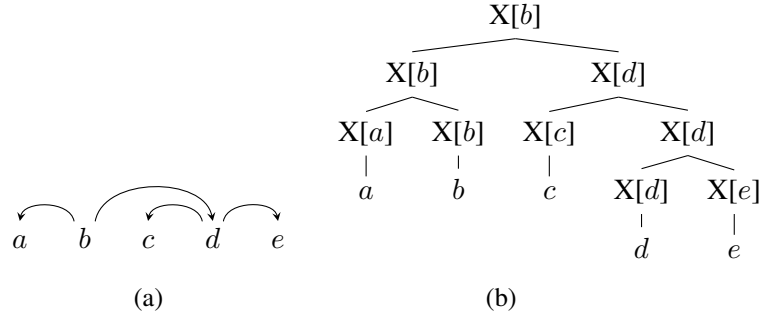


Figure 4.7: Implicit binarization process of the oracle described in the body.

- Given a subtree rooted at h , if the parent of it is its right side, or h is the sentence root, h 's left children are constructed first.
- If the parent of h is its left side, h 's right children are constructed first.

Figure 4.7 shows an example. For example, since the parent of d is b , which is in left, the constituent $d e$ is constructed first.

An important observation for showing this is the following lemma about the condition for applying RIGHTCOMP.

Lemma 4.1. *Let $c = (\sigma|\sigma_2|\sigma_1, \beta, A)$ and σ_2 be incomplete (next action is a reduce action). Then, in the above oracle, RIGHTCOMP never occurs for a configuration on which σ_2 is rooted at a dummy, i.e., $\sigma_2 = \langle x(\lambda) \rangle$, or σ_1 has some left children, i.e., $\exists k < \sigma_{11}, (\sigma_{11}, k)$.*

Proof. The first constraint, $\sigma_2 \neq \langle x(\lambda) \rangle$ is shown by simulating how a tree after RIGHTCOMP is created in the presented oracle. Let us assume $\lambda = \{i\}$ (i.e., σ_2 looks like $i \frown x$), $j = \sigma_{11}$, and σ_1 be a subtree spanning from $i+1$ to k (i.e., $i+1 \leq j \leq k$). After RIGHTCOMP, we get a subtree rooted at j , which looks like $i \frown j \frown x'$ where x' is a new dummy node. The oracle instead builds the same structure by the following procedures: After building $i \frown x$ by LEFTPRED to i , it first collect all left children of x by consecutive LEFTCOMP actions, followed by INSERT, to obtain a tree $i \frown j$ (omit j 's left dependents between i and j). Then it collects the right children of j (corresponding to x') with RIGHTPREDS. This is because we prefer LEFTCOMP and INSERT over LEFTPRED and SHIFT, and suggests that $\sigma_2 \neq \langle x(\lambda) \rangle$ before RIGHTCOMP. This simulation also implies the second constraint that $\nexists k < \sigma_{11}, (\sigma_{11}, k)$, since it never occurs unless LEFTPRED is preferred over LEFTCOMP. ■

Proof of Theorem 4.1. Let us examine the first case in which the parent of h is in the right side. Let this parent index be h' , i.e., $h < h'$. Note that this right-to-left arc ($h \frown h'$) are only created with LEFTCOMP or LEFTPRED and in both cases h must finish collecting every child before reduced, meaning that h' does not affect the construction of a subtree rooted at h . This is also the case when h is the sentence root. Now, if h collects its right children first, that means h collects left children via RIGHTCOMP with subtree rooted at a dummy node (which is later identified to h) but this never occurs by Lemma 4.1.

In the second case, $h' < h$. The arc $h' \frown h$ is established with **RIGHTPRED** or **RIGHTCOMP** when the head of the top stack symbol is h' (instantiated dummy node is later filled with h). In both cases, if h collects its left children first, that means a subtree rooted at h (the top stack symbol) with left children is substituted to the dummy node with **RIGHTCOMP** (see Figure 4.5). However, this situation is prohibited by Lemma 4.1. The oracle instead collects left children of h with successive **LEFTCOMPS**. This occurs due to the oracle's preference for **LEFTCOMP** over **LEFTPRED**. ■

Other notes

- The property of binarization above also indicates that the designed oracle is optimal in terms of stack depth, i.e., it always minimizes the maximum stack depth for a dependency tree, since it will minimize the number of turning points of the zig-zag path.
- If we construct another oracle algorithm, we would have different properties regarding implicit binarization, in which case Lemma 4.1 would not be satisfied.
- Combining the result in Section 4.3.2 about the transparency between the stack depth of the transition system and the left-corner PDA, it is obvious that at each step of this oracle, the incurred stack depth to recognize a dependency tree equals to the stack depth incurred by the left-corner PDA during recognizing the CFG parse given by the presented binarization.
- In Section 4.4, we use this oracle to evaluate the ability of the left-corner parser to recognize natural language sentences within small stack depth bound. Note if our interest is just to examine rarity of center-embedded constructions, that is possible without running the oracle in entire sentences, by just counting the degree of center-embedding of the binarized CFG parses. The main reason why we do not employ such method is because our interests are not limited to rarity of center-embedded constructions but also lie in the relative performance of the left-corner parser to capture syntactic regularities among other transition-based parsers, such as the arc-eager parser. This comparison seems more meaningful when we run every transition system on the same sentences. To make this comparison clearer, we next give more detailed analysis on the stack depth behavior of our left-corner transition system.

4.3.4 Stack depth of the transition system

We finally summarize the property of the left-corner transition system in terms of the stack depth. To do so, let us first introduce two measure, depth_{re} and depth_{sh} , with the former representing the stack depth after a reduce step and the latter representing the stack depth after a shift step. Then, we have:

- $\text{Depth}_{re} \leq 1$ unless the implicit CFG parse does not contain center-embedding (i.e., is just left-linear or right-linear). This linearly increases as the degree of center-embedding increases.
- $\text{Depth}_{sh} \leq 2$ if the implicit CFG parse does not contain center-embedding. The extra element on the stack occurs with a **SHIFT** action, but it does not imply the existence of center-embedding. This linearly increases as the degree of center-embedding increases.

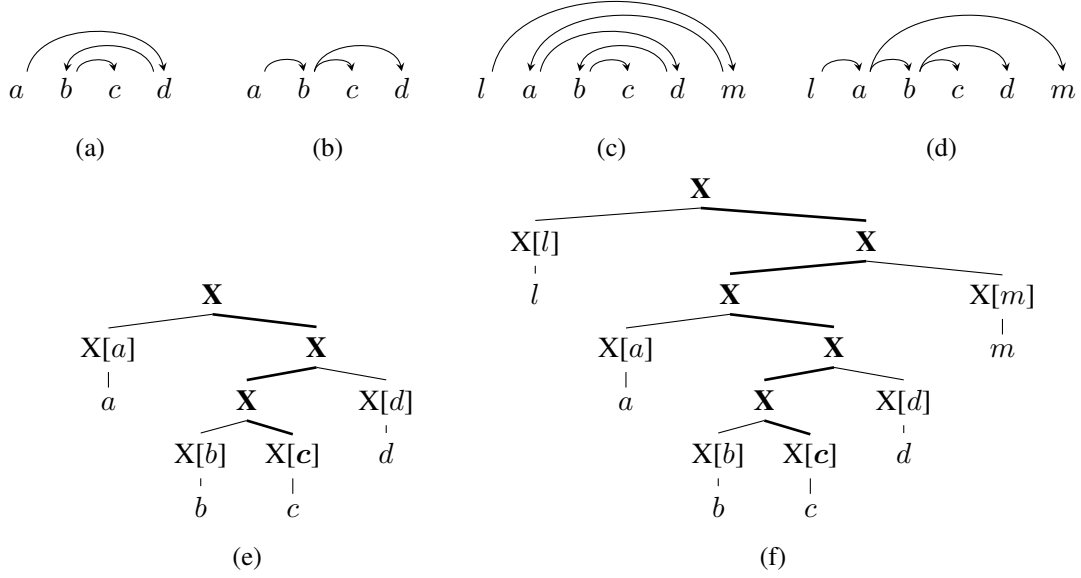


Figure 4.8: Center-embedded dependency trees and zig-zag patterns observed in the implicit CFG parses: (a)–(b) depth one, (c)–(d) depth two, (e) CFG parse for (a) and (b), and (f) CFG parse for (c) and (d).

The first statement about depth_{re} directly comes from Theorem 2.1 for the left-corner PDA. The second statement is about depth_{sh} , which we did not touch for the PDA. Figure 4.8 shows examples of how the depth of center-embedding increases, with the distinguished zig-zag patterns in center-embedded structures shown in bold. Note that depth_{re} can capture the degree of center-embedding correctly, by $\max \text{depth}_{re} - 1$ (Theorem 2.1), while depth_{sh} may not; for example, for parsing a right-branching structure $a \frown b \frown c$, b must be SHIFTed (not inserted) before being reduced, resulting in $\text{depth}_{sh} = 2$. We do not precisely discuss the condition with which an extra factor of depth_{sh} occurs. Of importance here is that both depth_{re} and depth_{sh} increase as the depth of center-embedding in the implicit CFG parse increases, though they may differ only by a constant (just one).

4.4 Empirical Stack Depth Analysis

In this section, we evaluate the cross-linguistic coverage of our developed transition system. We compare our system with other systems by observing the required stack depth as we run oracle transitions for sentences on a set of typologically diverse languages. We thereby verify the hypothesis that our system consistently demands less stack depth across languages in comparison with other systems. Note that this claim is not obvious from our theoretical analysis (Table 4.1) since the stack depth of the arc-eager system is sometimes smaller than that of the left-corner system (e.g., a subset of center-embedding), which suggests that it may possibly provide a more meaningful measure for capturing the syntactic regularities of a language.

4.4.1 Settings

Datasets We use two kinds of multilingual corpora introduced in Chapter 3, CoNLL dataset and Universal dependencies (UD), both of which comprises of 19 treebanks. Below, the first part of analyses in Sections 4.4.2, 4.4.3, and 4.4.4 are performed on CoNLL dataset while the latter analyses in Sections 4.4.5 and 4.4.6 are based on UD.

Since all systems presented in this chapter cannot handle nonprojective structures (Nivre, 2008), we projectivize all nonprojective sentences using pseudo-projectivization (Nivre and Nilsson, 2005) implemented in the MaltParser (Nivre et al., 2007b) (see also Section 2.1.5). We expect that this modification does not substantially change the overall corpus statistics as nonprojective constructions are relatively rare (Nivre et al., 2007a). Some treebanks such as the Prague dependency treebanks (including Arabic and Czech) assume that a sentence comprises multiple independent clauses that are connected via a dummy root token. We place this dummy root node at the end of each sentence, because doing so does not change the behaviors for sentences with a single root token in all systems and improves the parsing accuracy of some systems such as arc-eager across languages as compared with the conventional approach in which the dummy token is placed only at the beginning of each sentence (Ballesteros and Nivre, 2013).

Method We compare three transition systems: arc-standard, arc-eager, and left-corner. For each system, we perform oracle transitions for all sentences and languages, measuring stack depth at each configuration. The arc-eager system sometimes creates a subtree at the beginning of a buffer, in which case we increment stack depth by one.

Oracle We run an oracle transition for each sentence with each system. For the left-corner system, we implemented the algorithm presented in Section 4.3.3. For the arc-standard and arc-eager systems, we implemented oracles preferring reduce actions over shift actions, which minimizes the maximum stack depth.

4.4.2 Stack depth for general sentences

For each language in CoNLL dataset, we count the number of configurations of a specific stack depth while performing oracles on all sentences. Figure 4.9 shows the cumulative frequencies of configurations as the stack depth increases for the arc-standard, arc-eager, and left-corner systems. The data answer the question as to which stack depth is required to cover X% of configurations when recovering all gold trees. Note that comparing absolute values here is less meaningful since the minimal stack depth to construct an arc is different for each system, e.g., the arc-standard system requires at least two items on the stack, while the arc-eager system can create a right arc if the stack contains one element. Instead, we focus on the universality of each system’s behavior for different languages.

As discussed in Section 4.2.1, the arc-standard system can only process left-branching structures within a constant stack depth; such structures are typical in head-final languages such as Japanese or Turkish, and we observe this tendency in the data. The system performs differently in other languages, so the behavior is not consistent across languages.

The arc-eager and left-corner systems behave similarly for many languages, but we observe that there are some languages for which the left-corner system behaves similarly across numerous languages, while the arc-eager system tends to incur a larger stack depth. In particular, except Arabic, the left-corner system covers over 90% (specifically, over 98%) of configurations with a stack depth ≤ 3 . The arc-eager system also has 90% coverage in many languages with a stack depth ≤ 3 , though some exceptions exist, e.g., German, Hungarian, Japanese, Slovene, and Turkish.

We observe that results for Arabic are notably different from other languages. We suspect that this is because the average length of each sentence is very long (i.e., 39.3 words; see Table 4.2 for overall corpus statistics). Buchholz and Marsi (2006) noted that the parse unit of the Arabic treebank is not a sentence but a paragraph in which every sentence is combined via a dummy root node. To remedy this inconsistency of annotation units, we prepared the modified treebank, which we denote as Arabic* in the figure, by treating each child tree of the root node as a new sentence.⁶ The results then are closer to other language treebanks, especially Danish, which indicates that the exceptional behavior of Arabic largely originates with the annotation variety. From this point, we review the results of Arabic* instead of the original Arabic treebank.

4.4.3 Comparing with randomized sentences

The next question we examine is whether the observation from the last experiment, i.e., that the left-corner parser consistently demands less stack depth, holds only for naturally occurring or grammatically correct sentences. We attempt to answer this question by comparing oracle transitions on original treebank sentences and on (probably) grammatically incorrect sentences. We create these incorrect sentences using the method proposed by Gildea and Temperley (2007). We reorder words in each sentence by first extracting a directed graph from the dependency tree, and then randomly reorder the children of each node while preserving projectivity. Following Gildea and Temperley (2007), we remove punctuation from all corpora in this experiment beforehand, since how punctuation is attached to words is not essential.

The dotted lines shown in Figure 4.10 denote the results of randomized sentences for each system. There are notable differences in required stack depth between original and random sentences in many languages. For example, with a stack depth ≤ 3 , the left-corner system cannot reach 90% of configurations in many randomized treebanks such as Arabic*, Catalan, Danish, English, Greek, Italian, Portuguese, and Spanish. These results suggest that our system demands less stack depth only for naturally occurring sentences. For Chinese and Hungarian, the differences are subtle; however, the differences are also small for the other systems, which implies that these corpora have biases on graphs to reduce the differences.

4.4.4 Token-level and sentence-level coverage results

As noted in Section 4.3.4, the stack depth of the left-corner system in our experiments thus far is not the exact measurement of the degree of center-embedding of the construction due to an extra factor introduced by the SHIFT action. In this section, we focus on depth_{re} , which matches the degree of center-embedding and may be more applicable to some applications.

⁶We removed the resulting sentence if the length was one.

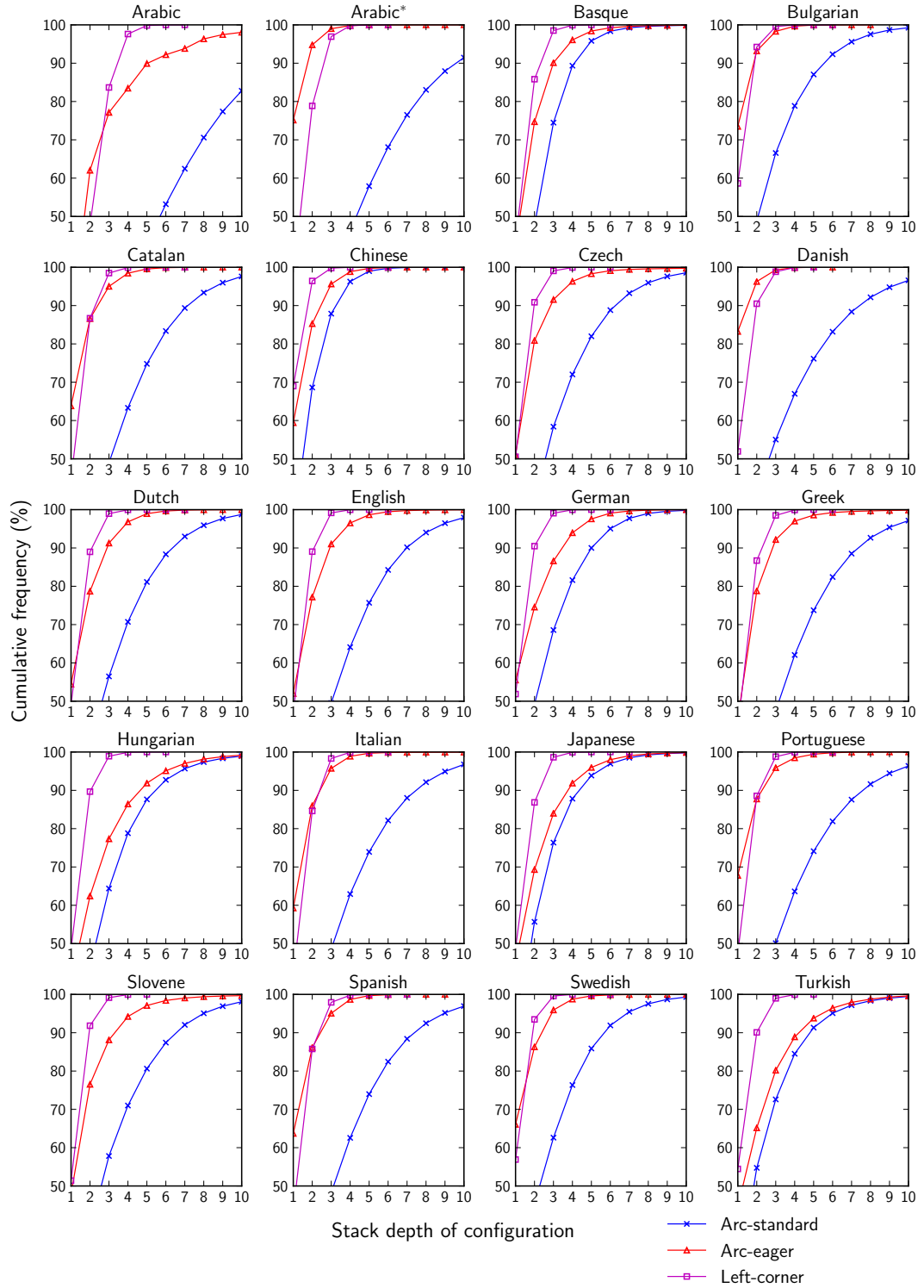


Figure 4.9: Crosslinguistic comparison of the cumulative frequencies of stack depth during oracle transitions.

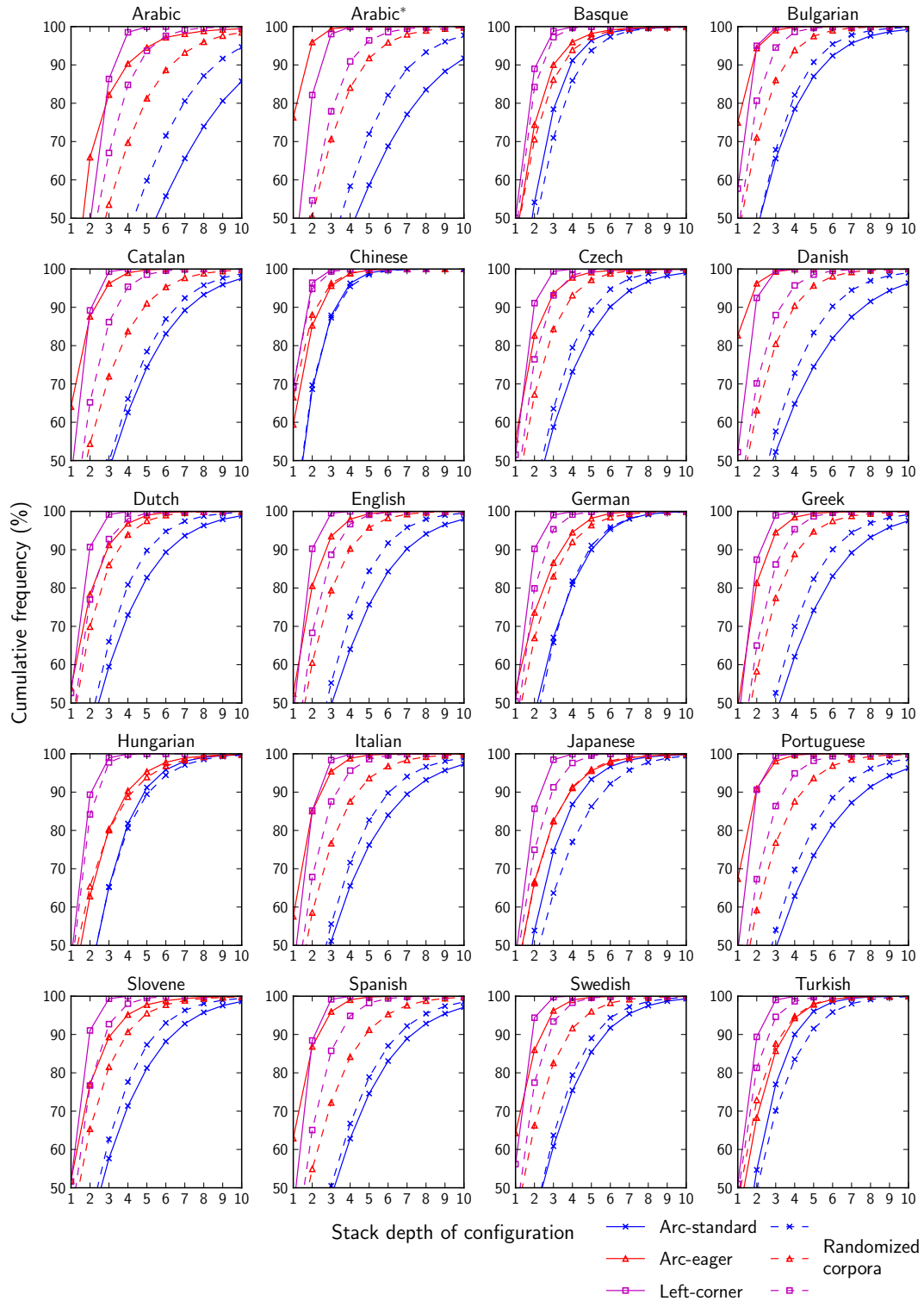


Figure 4.10: Stack depth results in corpora with punctuation removed; the dashed lines show results on randomly reordered sentences.

| | | Arabic | Arabic* | Basque | Bulgarian | Catalan | Chinese | Czech |
|----------|----------|-----------------|-------------------|----------------|------------------|----------------|------------------|----------------|
| #Sents. | | 3,043 | 4,102 | 3,523 | 13,221 | 15,125 | 57,647 | 25,650 |
| Av. len. | | 39.3 | 28.0 | 16.8 | 15.8 | 29.8 | 6.9 | 18.0 |
| Token | ≤ 1 | 22.9/21.8 | 52.8/55.9 | 57.3/62.7 | 79.5/80.0 | 66.2/69.4 | 83.6/83.6 | 74.7/74.0 |
| | ≤ 2 | 63.1/65.4 | 89.6/92.2 | 92.1/93.3 | 98.1/98.7 | 94.8/96.8 | 98.3/98.3 | 96.6/97.3 |
| | ≤ 3 | 92.0/94.1 | 98.9/99.4 | 99.2/99.3 | 99.8/99.9 | 99.5/99.8 | 99.9/99.9 | 99.7/99.8 |
| | ≤ 4 | 99.1/99.5 | 99.9/99.9 | 99.9/99.9 | 99.9/99.9 | 99.9/99.9 | 99.9/99.9 | 99.9/99.9 |
| Sent. | ≤ 1 | 7.0/7.4 | 20.8/21.4 | 15.5/20.8 | 37.3/39.4 | 14.7/16.9 | 58.3/58.3 | 32.0/34.2 |
| | ≤ 2 | 26.8/27.8 | 55.4/59.3 | 69.8/75.8 | 90.7/93.0 | 68.3/75.5 | 95.0/95.0 | 83.9/86.6 |
| | ≤ 3 | 57.6/61.6 | 91.7/94.5 | 95.8/97.0 | 99.3/99.7 | 95.6/98.3 | 99.7/99.7 | 98.2/99.1 |
| | ≤ 4 | 90.9/94.4 | 99.5/99.8 | 99.7/99.8 | 99.9/99.9 | 99.7/99.9 | 99.9/99.9 | 99.8/99.9 |
| | | Danish | Dutch | English | German | Greek | Hungarian | Italian |
| #Sents. | | 5,512 | 13,735 | 18,791 | 39,573 | 2,902 | 6,424 | 3,359 |
| Av. len. | | 19.1 | 15.6 | 25.0 | 18.8 | 25.1 | 22.6 | 23.7 |
| Token | ≤ 1 | 71.3/75.2 | 70.2/73.4 | 69.2/71.3 | 66.9/66.7 | 66.7/66.8 | 65.6/64.1 | 62.8/64.1 |
| | ≤ 2 | 95.6/97.4 | 95.9/96.8 | 96.3/97.5 | 94.5/94.5 | 95.2/96.2 | 95.1/94.9 | 94.0/94.2 |
| | ≤ 3 | 99.6/99.8 | 99.7/99.8 | 99.7/99.9 | 99.5/99.5 | 99.6/99.8 | 99.5/99.5 | 99.5/99.5 |
| | ≤ 4 | 99.9/99.9 | 99.9/99.9 | 99.9/99.9 | 99.9/99.9 | 99.9/100 | 99.9/99.9 | 99.9/99.9 |
| Sent. | ≤ 1 | 26.1/29.7 | 33.0/37.3 | 13.5/16.7 | 22.7/23.7 | 20.7/22.5 | 14.0/14.7 | 25.0/27.3 |
| | ≤ 2 | 77.9/83.4 | 83.4/87.3 | 73.4/80.0 | 71.3/72.8 | 76.6/80.8 | 69.3/70.4 | 76.0/77.2 |
| | ≤ 3 | 96.8/98.9 | 98.2/98.7 | 97.8/99.0 | 96.3/96.6 | 97.4/98.4 | 95.8/96.2 | 97.3/97.5 |
| | ≤ 4 | 99.8/99.9 | 99.8/99.9 | 99.8/99.9 | 99.7/99.7 | 99.8/100 | 99.7/99.7 | 99.8/99.8 |
| | | Japanese | Portuguese | Slovene | Spanish | Swedish | Turkish | |
| #Sents. | | 17,753 | 9,359 | 1,936 | 3,512 | 11,431 | 5,935 | |
| Av. len. | | 9.8 | 23.7 | 19.1 | 28.0 | 18.2 | 12.7 | |
| Token | ≤ 1 | 57.1/55.0 | 68.7/73.0 | 76.4/74.9 | 64.0/67.0 | 78.5/80.1 | 65.8/62.7 | |
| | ≤ 2 | 90.6/89.5 | 95.5/97.5 | 97.1/97.3 | 93.4/96.1 | 98.1/98.6 | 93.9/93.7 | |
| | ≤ 3 | 99.1/99.0 | 99.6/99.9 | 99.7/99.8 | 99.1/99.8 | 99.9/99.9 | 99.4/99.5 | |
| | ≤ 4 | 99.9/99.9 | 99.9/99.9 | 99.9/100 | 99.9/99.9 | 99.9/99.9 | 99.9/99.9 | |
| Sent. | ≤ 1 | 57.3/58.1 | 27.1/30.8 | 34.0/40.1 | 17.8/20.2 | 32.0/34.4 | 37.6/38.8 | |
| | ≤ 2 | 81.8/81.8 | 78.7/85.1 | 85.7/88.5 | 66.1/73.5 | 87.8/90.3 | 80.1/81.0 | |
| | ≤ 3 | 97.0/97.1 | 97.4/99.1 | 98.3/99.0 | 94.5/97.9 | 99.1/99.6 | 97.1/97.5 | |
| | ≤ 4 | 99.8/99.8 | 99.8/99.9 | 99.9/100 | 99.2/99.9 | 99.9/99.9 | 99.8/99.8 | |

Table 4.2: Token-level and sentence-level coverage results of left-corner oracles with depth_{re} . Here, the right-hand numbers in each column are calculated from corpora that exclude all punctuation, e.g., 92% of tokens in Arabic are covered within a stack depth ≤ 3 , while the number increases to 94.1 when punctuation is removed. Further, 57.6% of sentences (61.6% without punctuation) can be parsed within a maximum depth_{re} of three, i.e., the maximum degree of center-embedding is at most two in 57.6% of sentences. Av. len. indicates the average number of words in a sentence.

Table 4.2 shows token- and sentence-level statistics with and without punctuations. The token-level coverage of $\text{depth} \leq 2$ substantially improves from the results shown in Figure 4.9 in many languages, consistently exceeding 90% except for Arabic*, which indicates that many configurations of a stack depth of two in previous experiments are due to the extra factor caused by the SHIFT action rather than the deeper center-embedded structures. Results showing that the token-level coverage reaches 99% in most languages with $\text{depth}_{re} \leq 3$ indicate that the constructions with the degree three of center-embedding occurs rarely in natural language sentences. Overall, sentence-level coverage results are slightly decreased, but they are still very high, notably 95% – 99% with $\text{depth}_{re} \leq 3$ for most languages.

4.4.5 Results on UD

In the following two sections, we move on to UD, in which annotation styles are more consistent across languages. Figure 4.11 shows the result of the same analysis as the comparison in Section 4.9 on CoNLL dataset (i.e., Figure 4.9). We do not observe substantial differences between CoNLL dataset and UD. Again, the left-corner system is the most consistent across languages. This result is interesting in that it indicates the stack depth constraint of the left-corner system is less affected by the choice of annotation styles, since the annotation of UD is consistently content head-based while that of CoNLL dataset is (although consistently is lower) mainly function head-based.⁷ We will see this tendency in more detail by analyzing token-based statistics based on depth_{re} below.

4.4.6 Relaxing the definition of center-embedding

The token-level analysis on CoNLL dataset in Section 4.4.4 (Table 4.2) reveals that in most languages $\text{depth}_{re} \leq 2$ is a sufficient condition to cover most constructions but there are often relatively large gaps between $\text{depth}_{re} \leq 1$ (i.e., no center-embedding) and $\text{depth}_{re} \leq 2$ (i.e., at most one degree of center-embedding). We explore in this section constraints that exist in the middle between these two. We do so by relaxing the definition of center-embedding that we discussed in Section 2.2.1.

Recall that in our definition of center-embedding (Definition 2.2), we check whether the length of the most embedded constituent is larger than one (i.e., $|x| \geq 2$ in Eq. 2.1). In other words, the minimal length of the most embedded constituent for center-embedded structures is *two* in this case. Here, we relax this condition; for example, if assume the minimal length of most embedded clause is *three*, we recognize some portion of singly center-embedded structures (by Definition 2.2), in which the size of embedded constituent is one or two, to be not center-embedded.

Due to the transparency between the stack depth and the degree of center-embedding, this can be achieved by not increasing depth_{re} when the size (number of tokens including the dummy node) of the top stack element does not exceed the threshold, which is one in default (thus no reduction occurs).

⁷ A theoretical analysis of the effect of the annotation style is interesting, but is beyond the scope of the current study. We only claim that substantial differences are not observed in the present empirical analysis. Generally speaking, two dependency representations based on content-head and function-head, do not lead to the identical CFG representation with binarization, but as the meanings that they encode are basically the same (with different notions of head) we expect that the resulting differences in CFG forms are not substantial.

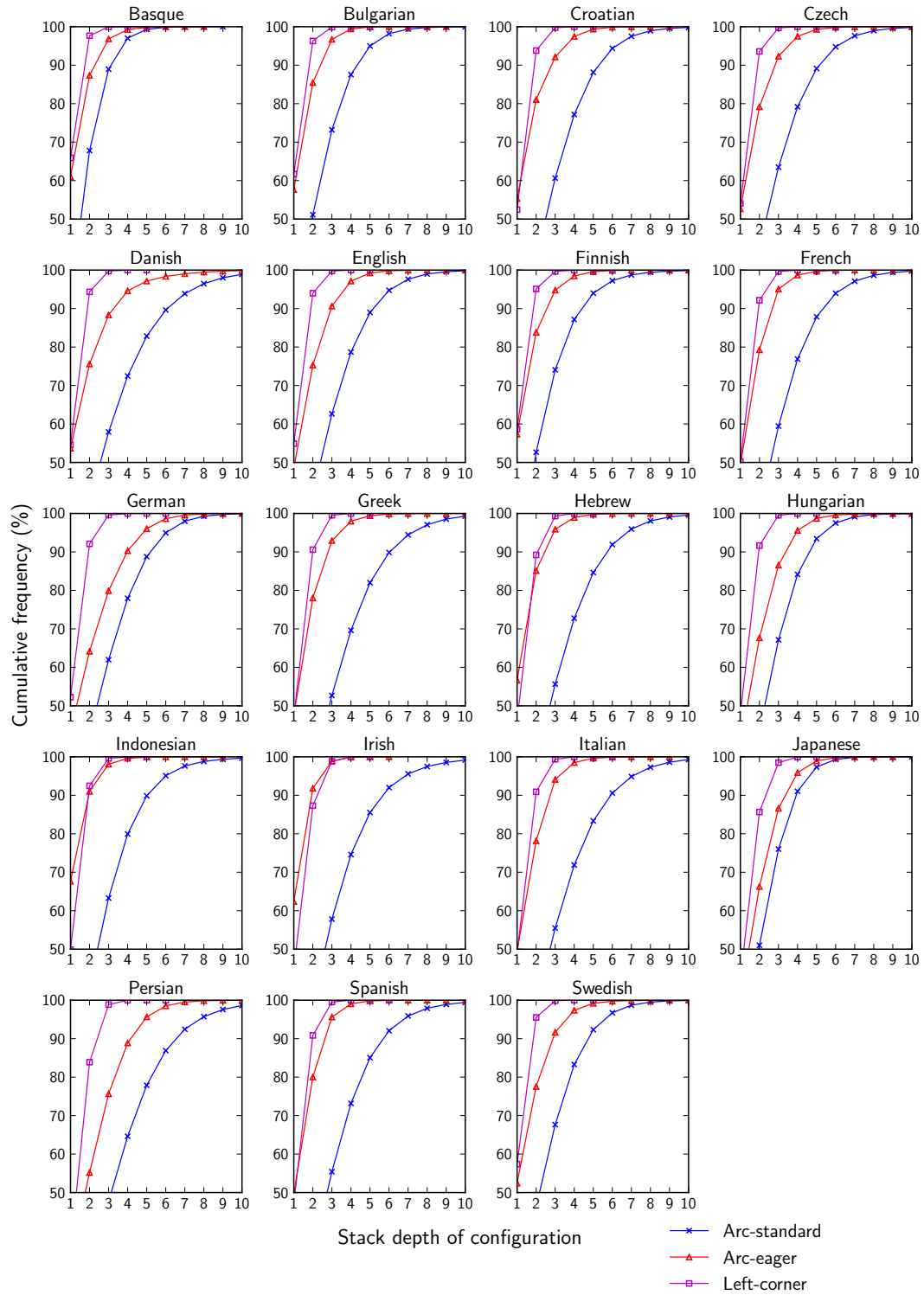


Figure 4.11: Stack depth results in UD.

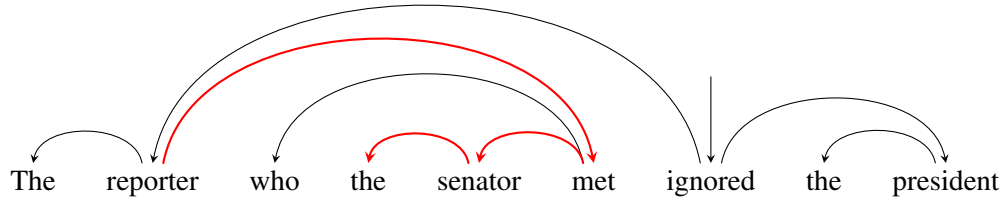


Figure 4.12: Following Definition 2.2, this tree is recognized as singly center-embedded while is not center-embedded if “the senator” is replaced by one word. Bold arcs are the cause of center-embedding (zig-zag pattern).

Motivating example In Section 2.2.6, we showed that the following sentence is recognized as not center-embedded when we follow Definition 2.2:

- (7) The reporter [who **Mary** met] ignored the president.

However, we can see that the following sentence is recognized as singly center-embedded:

- (8) The reporter [who **the senator** met] ignored the president.

Figure 4.12 shows the UD-style dependency tree with the emphasis on arcs causing center-embedding. This observation suggests many constructions that requires $\text{depth}_{re} = 2$ might be caught by relaxing the condition of center-embedding discussed above.

Result Figure 4.13 is the result with such relaxed conditions. Here we also show the effect of changing maximum sentence length. We can see in some languages, such as Hungarian, Japanese, and Persian, the effect of this relaxation is substantial while the changes in other languages are rather modest. We can also see that in most languages depth two is a sufficient condition to cover most constructions, which is again consistent with our observation in CoNLL dataset (Section 4.4.4).

We will explore this relaxation again in the supervised experiments we present below. Interestingly, there we will observe that the improvements with those relaxations are more substantial in parsing experiments (Section 4.5.5).

4.5 Parsing Experiment

Our final experiment is the parsing experiment on unseen sentences. A transition-based dependency parsing system is typically modeled with a structured discriminative model, such as with the structured perceptron and beam search (Zhang and Clark, 2008; Huang and Sagae, 2010). We implemented and trained the parser model in this framework to investigate the following questions:

- How does the stack depth bound at decoding affect parsing performance of each system? The underlying concern here is basically the same as in the previous oracle experiment discussed in Section 4.4, i.e., to determine whether the stack depth of the left-corner system provides

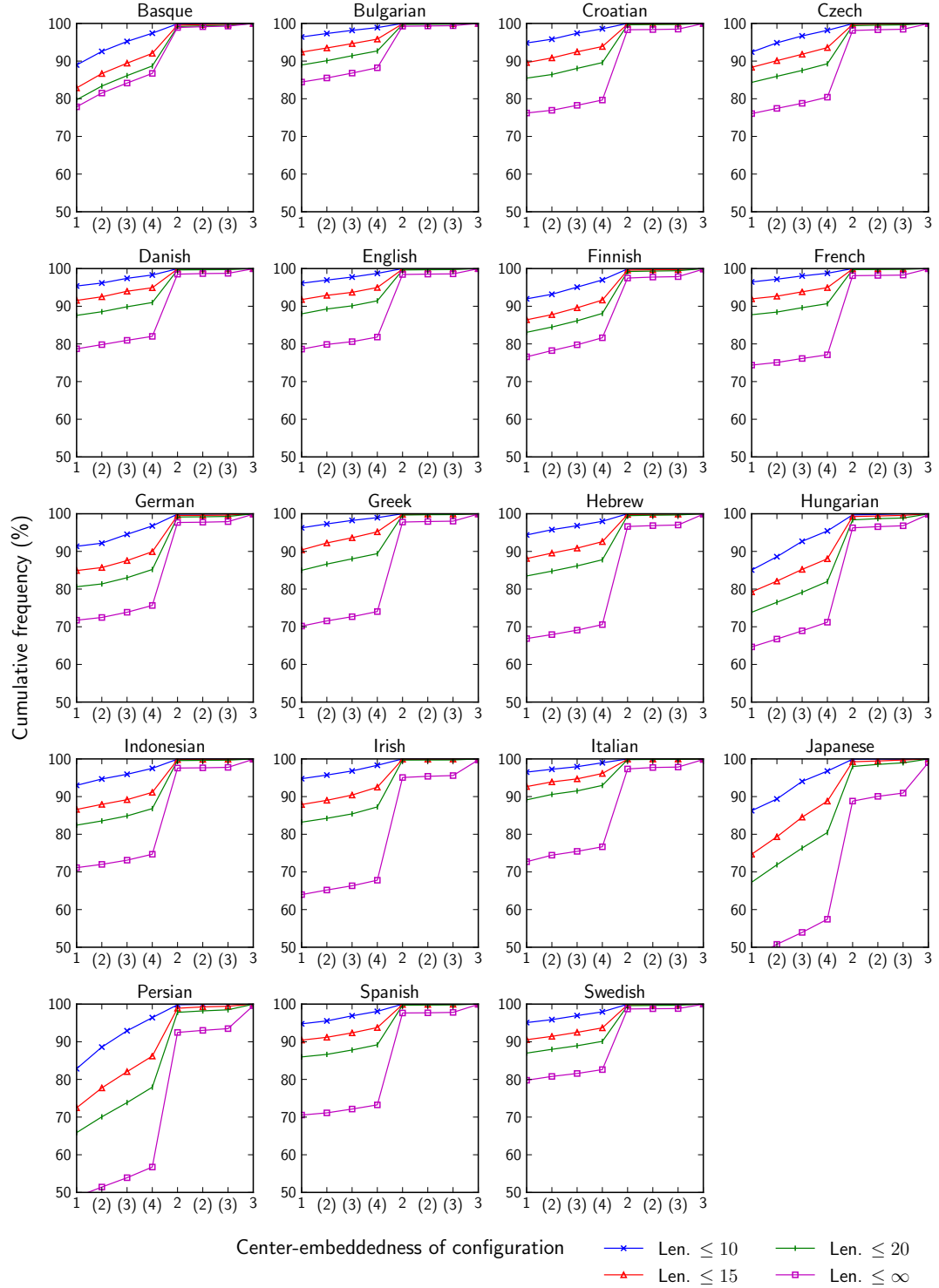


Figure 4.13: Stack depth results in UD with a left-corner system (depth_{re}) when the definition of center-embedding is relaxed. The parenthesized numbers indicate the size of allowed constituents at the bottom of embedding. For example (2) next to 2 indicates we allow depth = 3 if the size of subtree on the top of the stack is 1 or 2. Len. is the maximum sentence length.

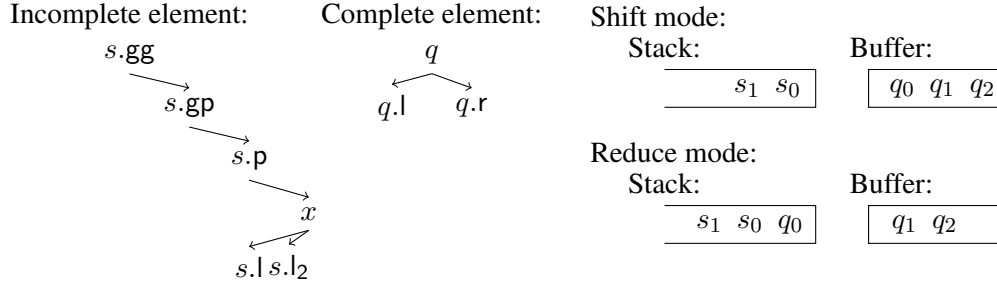


Figure 4.14: (Left) Elementary features extracted from an incomplete and complete node, and (Right) how feature extraction is changed depending on whether the next step is shift or reduce.

a meaningful measure for capturing the syntactic regularities. More specifically, we wish to observe whether the observation from the last experiment, i.e., that the behavior of the left-corner system is mostly consistent across languages, also holds with parse errors.

- Does our parser perform better than other transition-based parsers? One practical disadvantage of our system is that its attachment decisions are made more eagerly, i.e., that it has to commit to a particular structure at an earlier point; however, this also means the parser may utilize rich syntactic information as features that are not available in other systems. We investigate whether these rich features help disambiguation in practice.
- Finally, we examine parser performance of our system under a restriction on features to prohibit lookahead on the buffer. This restriction is motivated by the previous model of probabilistic left-corner parsing (Schuler et al., 2010) in which the central motivation is its cognitive plausibility. We report how accuracies drop with the cognitively motivated restriction and discuss a future direction to improve performance.

In the following we will investigate the above questions mainly with CoNLL dataset, as in our analysis in Section 4.4. In Section 4.5.2, we explain several experimental setups. We first compare the performances in the standard English experiments in Section 4.5.3, and then present experiments in CoNLL dataset in Section 4.5.4. Finally, we summarize the results in UD in Section 4.5.5.

4.5.1 Feature

The feature set we use is explained in Figure 4.14 and Tables 4.3 and 4.4. Our transition system is different from other systems in that it has two modes, i.e., a shift mode in which the next action is either SHIFT or INSERT and a reduce mode in which we select the next reduce action, thus we use different features depending on the current mode. Figure 4.14 shows how features are extracted from each node for each mode. In reduce mode, we treat the top node of the stack as if it were the top of buffer (q_0), which allows us to use the same feature templates in both modes by modifying only the definitions of elementary features s_i and q_i . A similar technique has been employed in the transition system proposed by Sartorio et al. (2013).

| | | | |
|---|--|---|---------------------------------------|
| $s_0.p.w$ | $s_0.p.t$ | $s_0.l.w$ | $s_0.l.t$ |
| $s_1.p.w$ | $s_1.p.t$ | $s_1.l.w$ | $s_1.l.t$ |
| $s_0.p.w \circ s_0.p.t$ | $s_0.l.w \circ s_0.l.t$ | $s_1.p.w \circ s_1.p.t$ | $s_1.l.w \circ s_1.l.t$ |
| $q_0.w$ | $q_0.t$ | $q_0.w \circ q_0.t$ | |
| $s_0.p.w \circ s_0.l.w$ | $s_0.p.t \circ s_0.l.t$ | | |
| $s_0.p.w \circ s_1.p.w$ | $s_0.l.w \circ s_1.l.w$ | $s_0.p.t \circ s_1.p.t$ | $s_0.l.t \circ s_1.l.t$ |
| $s_0.p.w \circ q_0.w$ | $s_0.l.w \circ q_0.w$ | $s_0.p.t \circ q_0.t$ | $s_0.l.t \circ q_0.t$ |
| $s_0.p.w \circ q_0.w \circ q_0.p$ | $s_0.p.w \circ q_0.w \circ s_0.p.t$ | $s_0.l.w \circ q_0.w \circ s_0.l.t$ | $s_0.l.w \circ q_0.w \circ s_0.l.t$ |
| $s_0.p.w \circ s_0.p.t \circ q_0.t$ | $s_0.l.w \circ s_0.l.t \circ q_0.t$ | | |
| $q_0.t \circ q_0.l.t \circ q_0.r.t$ | $q_0.w \circ q_0.l.t \circ q_0.r.t$ | | |
| $s_0.p.t \circ s_0.gp.t \circ s_0.gg.t$ | $s_0.p.t \circ s_0.gp.t \circ s_0.l.t$ | $s_0.p.t \circ s_0.l.t \circ s_0.l_2.t$ | $s_0.p.t \circ s_0.gp.t \circ q_0.t$ |
| $s_0.p.t \circ s_0.l.t \circ q_0.t$ | $s_0.p.w \circ s_0.l.t \circ q_0.t$ | $s_0.p.t \circ s_0.l.w \circ q_0.t$ | $s_0.l.t \circ s_0.l_2.p \circ q_0.t$ |
| $s_0.l.t \circ s_0.l_2.t \circ q_0.t$ | $s_0.p.t \circ q_0.t \circ q_0.l.t$ | $s_0.p.t \circ q_0.t \circ q_0.r.t$ | |
| $s_1.p.t \circ s_0.p.t \circ s_0.l.t$ | $s_1.p.t \circ s_0.l.t \circ q_0.t$ | $s_1.l.t \circ s_0.l.t \circ q_0.t$ | $s_1.l.t \circ s_0.l.t \circ q_0.t$ |
| $s_1.l.t \circ s_0.p.t \circ q_0.p$ | | | |

Table 4.3: Feature templates used in both full and restricted feature sets, with t representing POS tag and w indicating the word form, e.g., $s_0.l.t$ refers to the POS tag of the leftmost child of s_0 . \circ means concatenation.

| | | | |
|-----------------------------------|-----------------------------------|---|---|
| $q_0.t \circ q_1.t$ | $q_0.t \circ q_1.t \circ q_2.t$ | $s_0.p.t \circ q_0.p \circ q_1.p \circ q_2.p$ | $s_0.l.t \circ q_0.t \circ q_1.t \circ q_2.t$ |
| $s_0.p.w \circ q_0.t \circ q_1.t$ | $s_0.p.t \circ q_0.t \circ q_1.t$ | $s_0.l.w \circ q_0.t \circ q_1.t$ | $s_0.l.t \circ q_0.t \circ q_1.t$ |

Table 4.4: Additional feature templates only used in the full feature model.

To explore the last question, we develop two feature sets. Our full feature set consists of features shown in Tables 4.3 and 4.4. For the limited feature set, we remove all features that depend on q_1 and q_2 in Figure 4.14, which we list in Table 4.4. Here, we only look at the top node on the buffer in shift mode. This is the minimal amount of lookahead in our parser and is the same as the previous left-corner PCFG parsers (Schuler et al., 2010), which are cognitively motivated.

Our parser cannot capture a head and dependent relationship directly at each reduce step, because all interactions between nodes are via a dummy node, which may be a severe limitation from a practical viewpoint; however, we can exploit richer context from each subtree on the stack, as illustrated in Figure 4.14. We construct our feature set with many nodes around the dummy node, including the parent (p), grandparent (gp), and great grandparent (gg).

4.5.2 Settings

We compare parsers with three transition systems: arc-standard, arc-eager, and left-corner. The feature set of the arc-standard system is borrowed from Huang and Sagae (2010). For the arc-eager system, we use the feature set of Zhang and Nivre (2011) from which we exclude features that rely on arc label information.

We train all models with different beam sizes in the violation fixing perceptron framework

(Huang et al., 2012). Since our goal is not to produce a state-of-the-art parsing system, we use gold POS tags as input both in training and testing.

As noted in Section 4.3.2, the left-corner parser sometimes fails to generate a single tree, in which case the stack contains a complete subtree at the top position (since the last action is always INSERT) and one or more incomplete subtrees. If this occurs, we perform the following post-processing steps:

- We collapse each dummy node in an incomplete tree. More specifically, if the dummy node is the head of the subtree, we attach all children to the sentence (dummy) root node; otherwise, the children are reattached to the parent of the dummy node.
- The resulting complete subtrees are all attached to the sentence (dummy) root node.

4.5.3 Results on the English Development Set

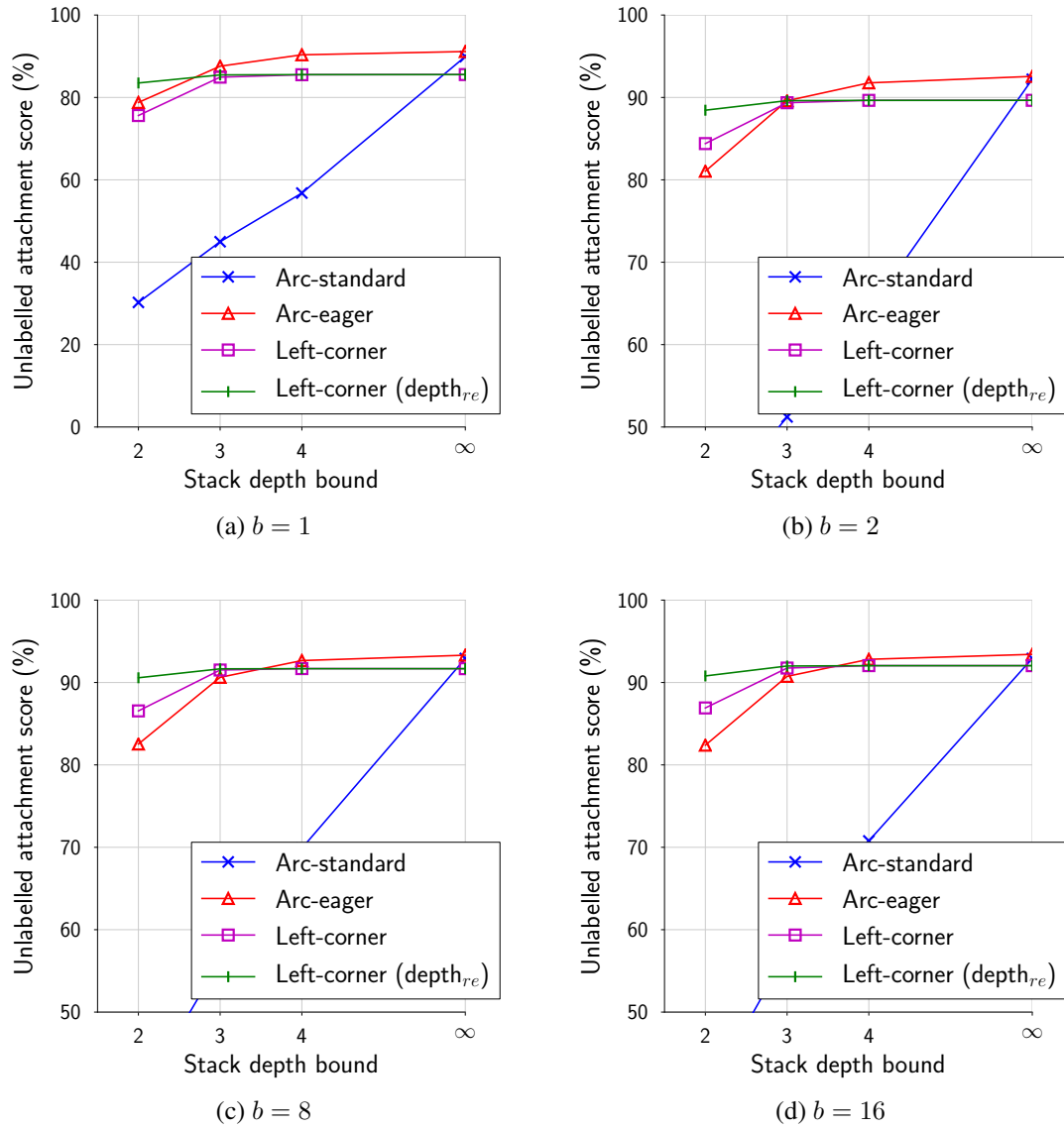
We first evaluate our system on the common English development experiment. We train the model in section 2-21 of the WSJ Penn Treebank (Marcus et al., 1993), which are converted into dependency trees using the LTH conversion tool⁸.

Impact of Stack Depth Bound To explore the first question posed at the beginning of this section, we compare parse accuracies under each stack depth bound with several beam sizes, with results shown in Figure 4.15. In this experiment, we calculate the stack depth of a configuration in the same way as our oracle experiment (see Section 4.4.1), and when expanding a beam, we discard candidates for which stack depth exceeds the maximum value. As discussed in Section 4.4.4, for the left-corner system, depth_{re} might be a more linguistically meaningful measure, so we report scores with both definitions.⁹ The general tendency across different beam sizes is that our left-corner parser (in particular with depth_{re}) is much less sensitive to the value of the stack depth bound. For example, when the beam size is eight, the accuracies of the left-corner (depth_{re}) are 90.6, 91.7, 91.7, and 91.7 with increased stack depth bounds, while the corresponding scores are 82.5, 90.6, 92.6, and 93.3 in the arc-eager system. This result is consistent with the observation in our oracle coverage experiment discussed in Section 4.4, and suggests that a depth bound of two or three might be a good constraint for restricting tree candidates for natural language with no (or little) loss of recall. Next, we examine whether this observation is consistent across languages.

Performance without Stack Depth Bound Figure 4.16 shows accuracies with no stack depth bound when changing beam sizes. We can see that the accuracy of the left-corner system (full feature) is close to that of the other two systems, but some gap remains. With a beam size of 16, the scores are left-corner: 92.0; arc-standard: 92.9; arc-eager: 93.4. Also, the score gaps are relatively large at smaller beam sizes; e.g., with beam size 1, the score of the left-corner is 85.5, while that of the arc-eager is 91.1. This result indicates that the prediction with our parser might be structurally harder than other parsers even though ours can utilize richer context from subtrees on the stack.

⁸http://nlp.cs.lth.se/software/treebank_converter/

⁹ This can be achieved by allowing any configurations after a shift step.

Figure 4.15: Accuracy vs. stack depth bound at decoding for several beam sizes (b).

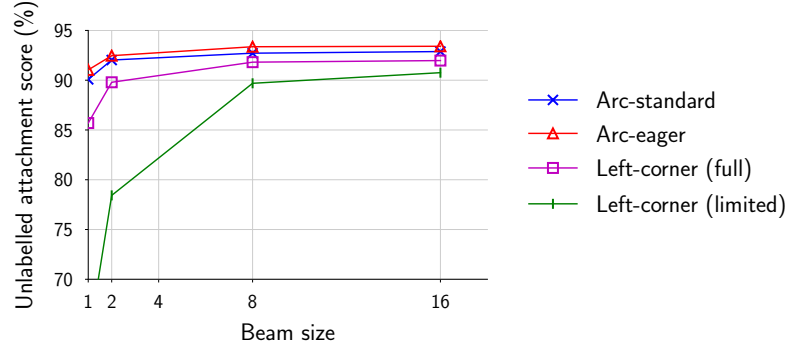


Figure 4.16: Accuracy vs. beam size for each system on the English Penn Treebank development set. Left-corner (full) is the model with the full feature set, while Left-corner (limited) is the model with the limited feature set.

Performance of Limited Feature Model Next we move on to the results with cognitively motivated limited features (Figure 4.16). When the beam size is small, it performs extremely poorly (63.6% with beam size 1). This is not surprising since our parser has to deal with each attachment decision much earlier, which seems hard without lookahead features or larger beam. However, it is interesting that it achieves a reasonably higher score of 90.6% accuracy with beam size 16. In the previous constituency left-corner parsing experiments that concerned their cognitive plausibility (Schuler et al., 2010; van Schijndel et al., 2013), typically the beam size is quite huge, e.g., 2,000. The largest difference between our parser and their systems is the model: our model is discriminative while their models are generative. Though discriminative models are not popular in the studies of human language processing (Keller, 2010), the fact that our parser is able to output high quality parses with such smaller beam size would be appealing from the cognitive viewpoint (see Section 4.6 for further discussion).

4.5.4 Result on CoNLL dataset

We next examine whether the observations above with English dataset are general across languages using CoNLL dataset. Note that although we train on the projectivized corpus, evaluation is against the original nonprojective trees. As our systems are unlabeled, we do not try any post-deprojectivization (Nivre and Nilsson, 2005). In this experiment, we set the beam size to 8.

Effect of Stack Depth Bound The cross-linguistic results with stack depth bounds are summarized in Figure 4.17 from which we can see that the overall tendency of each system is almost the same as the English experiment. Little accuracy drops are observed between models with bounded depth 2 or 3 and the model without depth bound in the left-corner (depth_{re}), although the score gaps are larger in the arc-eager. The arc-standard parser performs extremely poorly with small depth bounds except Japanese and Turkish, and this is consistent with our observation that the arc-standard system demands less stack depth only for head-final languages (Section 4.4.2).

Notably, in some cases the scores of the left-corner parser (depth_{re}) drop when loosening the depth

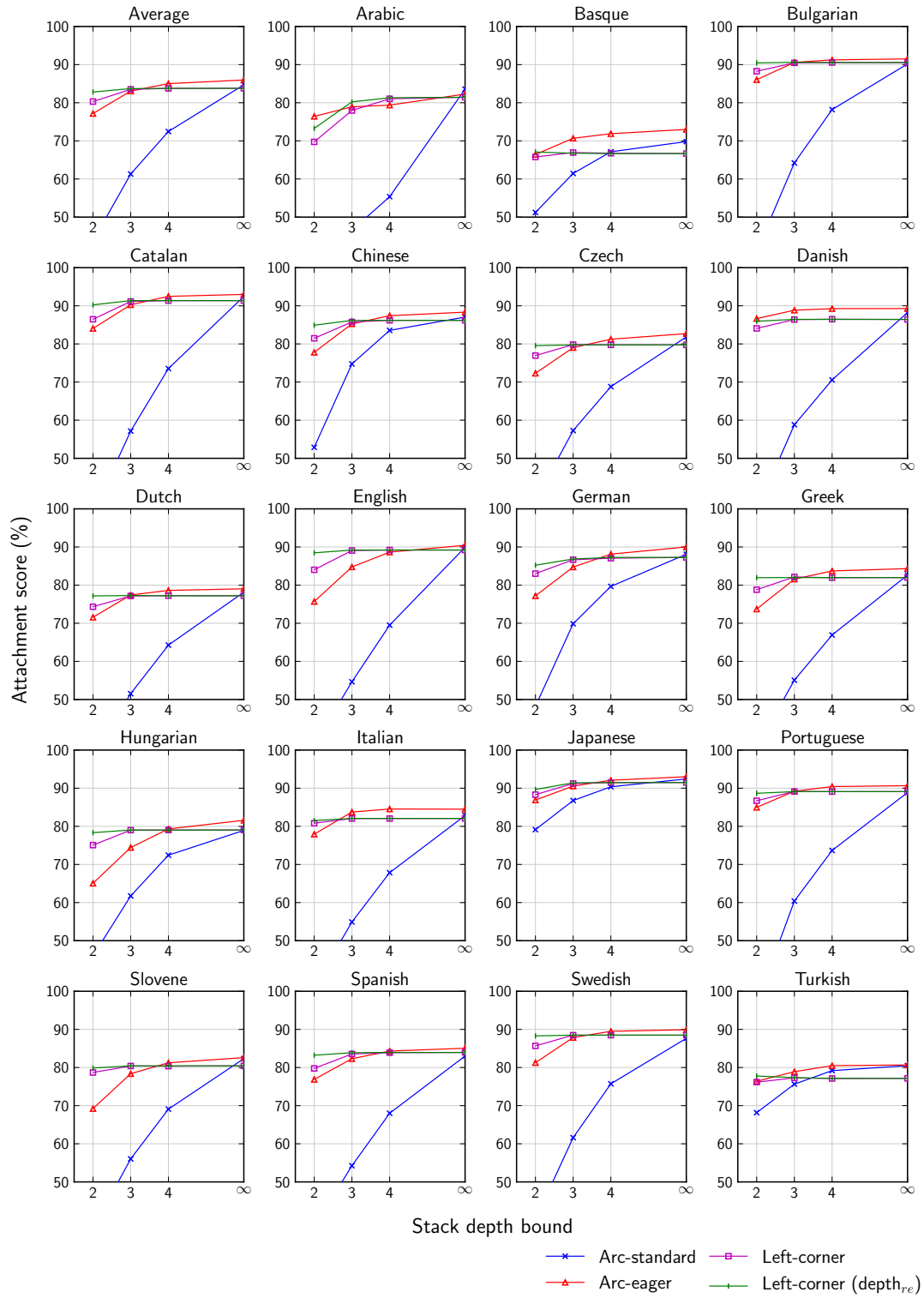


Figure 4.17: Accuracy vs. stack depth bound in CoNLL dataset.

| | Arc-standard | Arc-eager | Left-corner full | Left-corner limited |
|------------|--------------|-----------|---------------------|------------------------|
| Arabic | 83.9 | 82.2 | 81.2 | 77.5 |
| Basque | 70.5 | 72.8 | 66.8 | 64.6 |
| Bulgarian | 90.2 | 91.4 | 89.9 | 88.1 |
| Catalan | 92.5 | 93.3 | 91.4 | 89.3 |
| Chinese | 87.3 | 88.4 | 86.8 | 83.6 |
| Czech | 81.5 | 82.3 | 80.1 | 77.2 |
| Danish | 88.0 | 89.1 | 86.8 | 85.5 |
| Dutch | 77.7 | 79.0 | 77.4 | 74.9 |
| English | 89.6 | 90.3 | 89.0 | 85.8 |
| German | 88.1 | 90.0 | 87.2 | 85.7 |
| Greek | 82.2 | 84.0 | 82.0 | 80.7 |
| Hungarian | 79.1 | 80.9 | 79.0 | 75.8 |
| Italian | 82.3 | 84.8 | 81.7 | 79.4 |
| Japanese | 92.5 | 92.9 | 91.3 | 90.7 |
| Portuguese | 89.2 | 90.6 | 88.9 | 87.1 |
| Slovene | 82.3 | 82.3 | 80.8 | 77.1 |
| Spanish | 83.0 | 85.0 | 83.8 | 80.6 |
| Swedish | 87.2 | 90.0 | 88.5 | 87.0 |
| Turkish | 80.8 | 80.8 | 77.5 | 75.4 |
| Average | 84.6 | 85.8 | 83.7 | 81.4 |

Table 4.5: Parsing results on CoNLL X and 2007 test sets with no stack depth bound (unlabeled attachment scores).

bound (see Basque, Danish, and Turkish), meaning that the stack depth bound of the left-corner sometimes help disambiguation by ignoring linguistically implausible structures (deep center-embedding) during search. The result indicates the parser performance could be improved by utilizing stack depth information of the left-corner parser, though we leave further investigation as a future work.

Performance without Stack Depth Bound Table 4.5 summarizes the results without stack depth bounds. Again, the overall tendency is the same as the English experiment. The arc-eager performs the best except Arabic. In some languages (e.g., Bulgarian, English, Spanish, and Swedish), the left-corner (full) performs better than the arc-standard, while the average score is 1.1 point below. This difference and the average difference between the arc-eager and the arc-standard (85.8 vs. 84.6) are both statistically significant ($p < 0.01$, the McNemar test). We can thus conclude that our left-corner parser is not stronger than the other state-of-the-art parsers even with rich features.

Performance of Limited Feature Model With limited features, the left-corner parser performs worse in all languages. The average score is about 2 point below the full feature models (83.7% vs.

81.4%) and shows the same tendency as in the English development experiment. This difference is also statistically significant ($p < 0.01$, the McNemar test). The scores of English are relatively low compared with the results in Table 4.16, probably because the training data used in the CoNLL 2007 shared task is small, about half of our development experiment, to reduce the cost of training with large corpora for the shared task participants (Nivre et al., 2007a).

Finally, though the overall score of the left-corner parser is lower, we suspect that it could be improved by inventing new features, in particular those with external syntactic knowledge. The analysis below is based on the result with limited features, but we expect a similar technique would also be helpful to the full feature model.

As we have discussed (see the beginning of Section 4.5), an attachment decision of the left-corner parser is more eager, which is the main reason for the lower scores. One particular difficulty with the left-corner parser is that the parser has to decide whether each token has further (right) arguments with no (or a little) access to the actual right context. Figure 4.18 shows an example of a parse error in English caused by the left-corner parser with limited features (without stack depth bound). This is a kind of PP attachment error on *on CNN*, though the parser has to deal with this attachment decision implicitly before observing the attached phrase (*on CNN*). When the next token in the buffer is *times* (Figure 4.18(c)), performing SHIFT means *times* would take more than one argument in future, while performing INSERT means the opposite: *times* does not take any arguments. Resolving this problem would require knowledge on *times* that it often takes no right arguments (while *appear* generally takes several arguments), but it also suggests that the parser performance could be improved by augmenting such syntactic knowledge on each token as new features, such as with distributional clustering (Koo et al., 2008; Bohnet et al., 2013), supertagging (Ouchi et al., 2014), or refined POS tags (Mueller et al., 2014). All those features are shown to be effective in transition-based dependency parsing; we expect those are particularly useful for our parser though the further analysis is beyond the scope of this chapter. In PCFG left-corner parsing, van Schijndel et al. (2013) reported accuracy improvement with symbol refinements obtained by the Berkeley parser (Petrov et al., 2006) in English.

4.5.5 Result on UD

Figure 4.19 shows the results in UD. Again the performance tendency is not changed from the CoNLL dataset; on average, the left-corner with depth_{re} can parse sentences without dropping accuracies but other systems are largely affected by the constraints.

We further examine the behavior of the left-corner parser by relaxing the definition of center-embedding which we discussed in Section 4.4.6. Figure 4.20 shows the result when we change the definition of depth_{re} . It is interesting to see that compared to Figure 4.13, the number of languages in which this relaxation had a greater impact increases; e.g., in Croatian, Czech, Danish, Finnish, Hungarian, Indonesian, Persian, and Swedish, there is about 10% improvements from the original $\text{depth}_{re} \leq 1$ to the relaxed $\text{depth}_{re} 1(3)$ (i.e., when three word constituents are allowed to be embedded). The reason of this might be in the characteristics of the supervised parsers, which more freely explore the search space (compared to the statistic analysis in Figure 4.13).

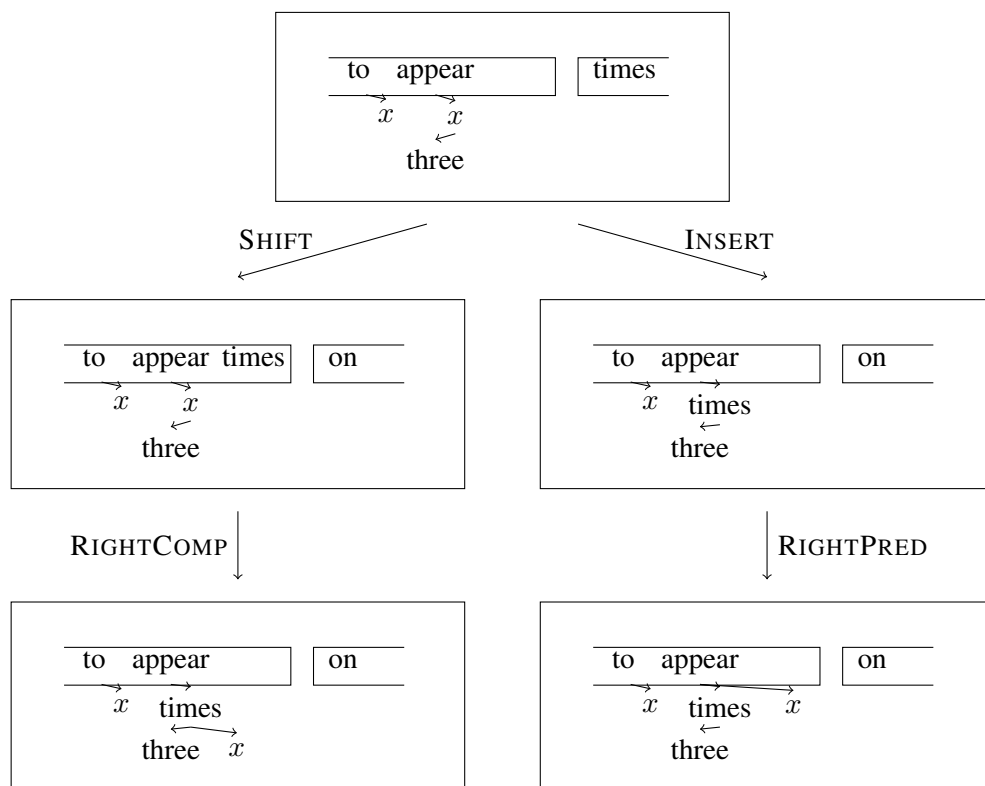
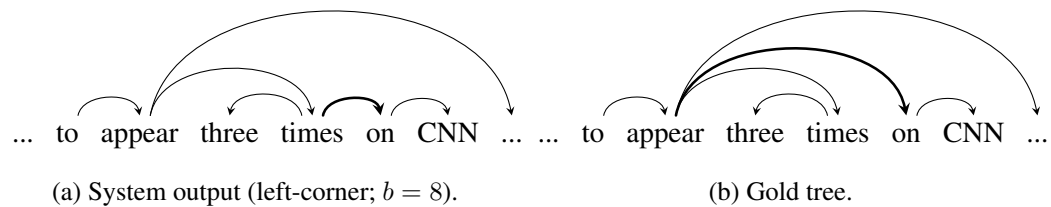


Figure 4.18: (a)-(b) Example of a parse error by the left-corner parser that may be saved with external syntactic knowledge (limited features and beam size 8). (c) Two corresponding configuration paths: the left path leads to (a) and the right path leads to (b).

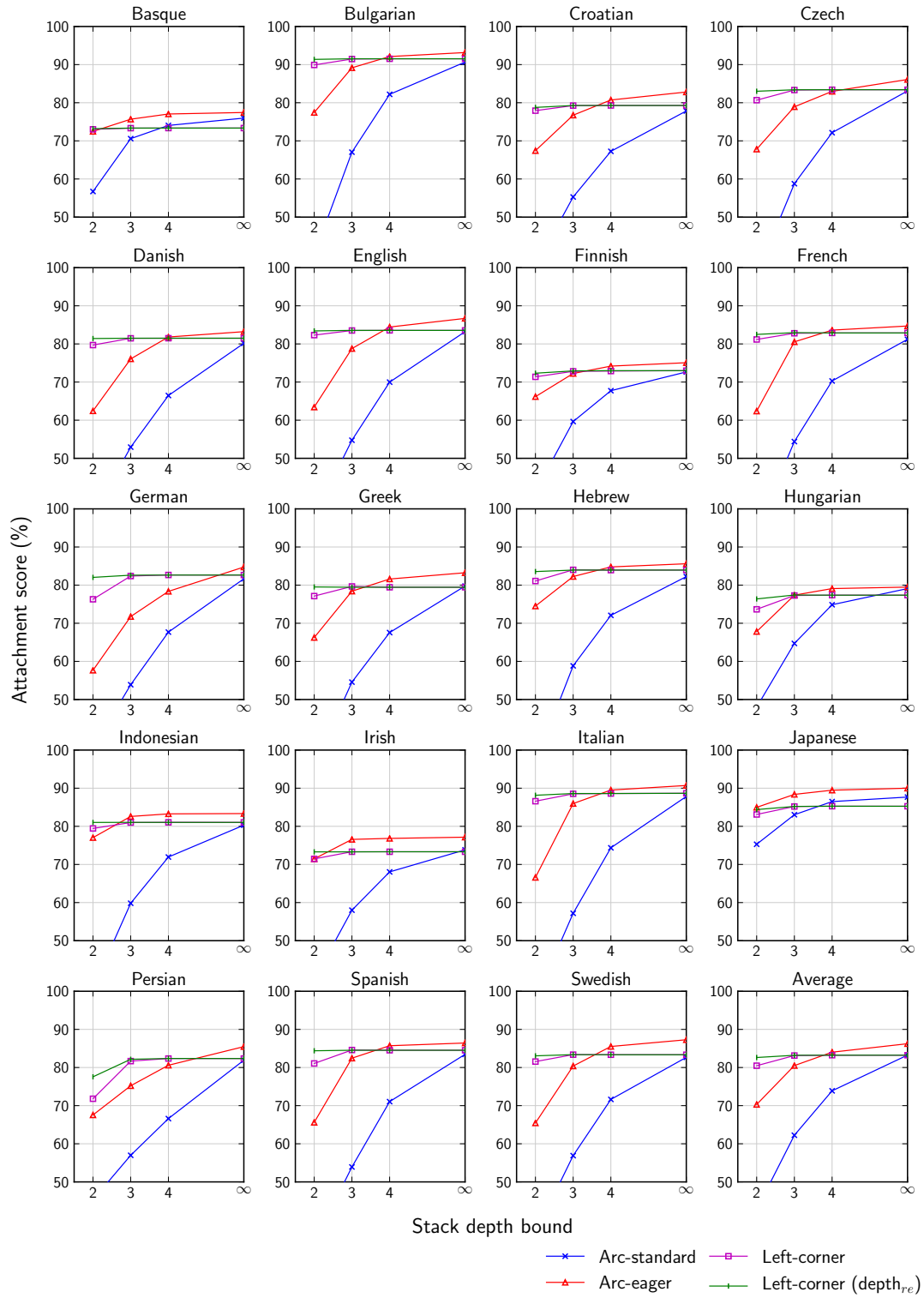


Figure 4.19: Accuracy vs. stack depth bound in UD.

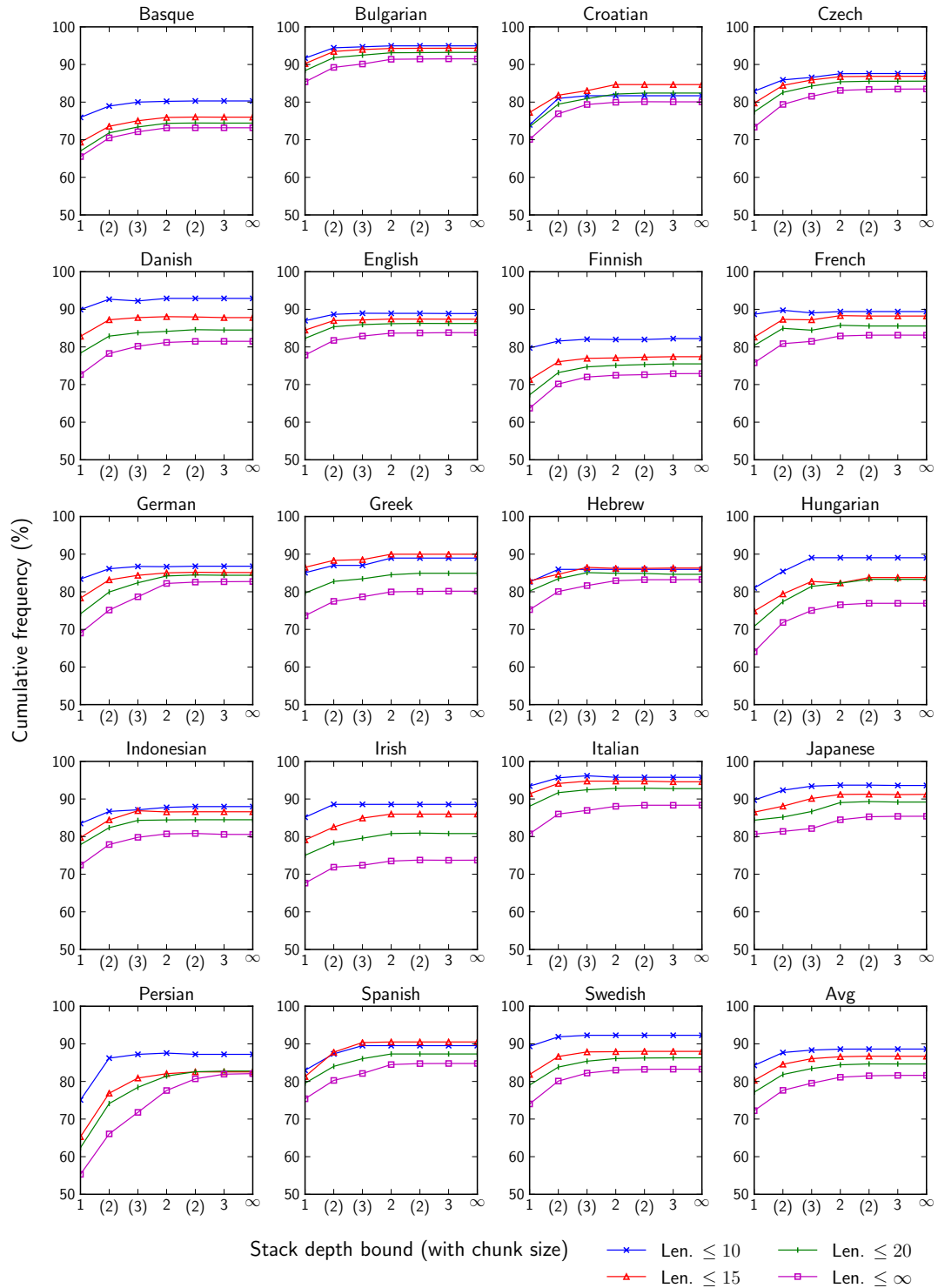


Figure 4.20: Accuracy vs. stack depth bound with left-corner parsers on UD with different maximum length of test sentences.

4.6 Discussion and Related Work

We have presented the left-corner parsing algorithm for dependency structures and showed that our parser demands less stack depth for recognizing most of natural language sentences. The result also indicates the existence of universal syntactic biases that center-embedded constructions are rare phenomena across languages. We finally discuss the relevance of the current study to the previous works.

We have reviewed previous works about left-corner parsing (for CFGs) in Section 2.2 though have little mentioned previous works that study the empirical property of the left-corner parsers. Roark (2001) is the first attempt of the empirical study. His idea is instead of modeling left-corner transitions directly as in our parser, incorporating the left-corner strategy into a CFG parser via a left-corner grammar transform (Johnson, 1998a). This design makes the overall parsing system top-down and makes it possible to compare the pure top-down and the left-corner parsing systems in a unified way. Note also that as his method is based on Johnson (1998a), the parsing mechanism is basically the same as the left-corner PDA that we introduced as another variant in Section 2.2.5. Schuler et al. (2010) examine the empirical coverage result of the left-corner PDA that we formalized in Section 2.2.3, though the experiment is limited on English.

Most of previous left-corner parsing models have been motivated by the study of cognitively plausible parsing models, an interdisciplinary research on psycholinguistics and computational linguistics (Keller, 2010). Though we also evaluated our parser with cognitively motivated limited feature models and got an encouraging result, this is preliminary and we do not claim from this experiment that our parser is cross-linguistically cognitively plausible. Our parser is able to parse most sentences within a limited stack depth bound. However, it is skeptical whether there is any connection between the stack of our parser and memory units preserved in human memory. van Schijndel and Schuler (2013) calculated several kinds of *memory cost* obtained from a configuration of their left-corner parser and discussed which cost is more significant indicator to predict human reading time data, such as the current stack depth and the integration cost in the dependency locality theory (Gibson, 2000), which is obtained by calculating the distance between two subtrees at composition. Discussing cognitive plausibility of a parser requires such kind of careful experimental setup, which is beyond the scope of the current work.

Our main focus in this chapter is rather a syntactic bias exist in language universally. In this view, our work is more relevant to previous dependency parsing model with a constraint on possible tree structures (Eisner and Smith, 2010). They studied parsing with a hard constraint on dependency length based on the observation that grammar may favor a construction with shorter dependency lengths (Gildea and Temperley, 2007; Gildea and Temperley, 2010). Instead of prohibiting longer dependency lengths, our method prohibits deeper center-embedded structures, and we have shown that this bias is effective to restrict natural language grammar. The two constraints, length and center-embedding, are often correlated since center-embedding constructions typically lead to longer dependency length. It is therefore an interesting future topic to explore which bias is more essential for restricting grammar. This question can be perhaps explored through unsupervised dependency parsing tasks (Klein and Manning, 2004), where such kind of light supervision has significant impact on the performance (Smith and Eisner, 2006; Mareček and Žabokrtský, 2012; Bisk and Hockenmaier, 2013).

We introduced a dummy node for representing a subtree with an unknown head or dependent. Recently, Menzel and colleagues (Beuck and Menzel, 2013; Kohn and Menzel, 2014) have also studied dependency parsing with a dummy node. While conceptually similar, the aim of introducing a dummy node is different between our approach and theirs: We need a dummy node to represent a subtree corresponding to that in Resnik’s algorithm, while they introduced it to confirm that every dependency tree on a sentence prefix is fully connected. This difference leads to a technical difference; a subtree of their parser can contain more than one dummy node, while we restrict each subtree to containing only one dummy node on a right spine.

Chapter 5

Grammar Induction with Structural Constraints

In the previous chapter, we formulated a left-corner dependency parsing algorithm as a transition system in which its stack size grows only for center-embedded constructions. Also, we investigated how much the developed parser can capture the syntactic biases found in the manually developed treebanks, and found that very restricted stack depth such as two or one (by allowing small constituents to be embedded) suffices to describe most syntactic constructions across languages.

In this chapter, we will investigate whether the found syntactic bias in the previous chapter would be helpful for the task of *unsupervised grammar induction*, where the goal is to learn the model of finding hidden syntactic structures given the surface strings (or part-of-speeches) alone; see Section 2.4 for overviews.

There are a number of motivations to consider unsupervised grammar induction, in particular with the *universal* syntactic biases as we discussed in Chapter 1. Among them our primary motivation is to investigate a good *prior* that would be useful for restricting possible tree structures for general natural language sentences (regardless of language). The structure that we aim to induce is dependency structure; though this choice mainly stems from computational reasons rather than philosophical ones, i.e., the dependency structure is currently the most feasible structure to be learned, we argue the lesson from the current study would be useful for the problem of inducing other structures including constituent-based representations, e.g., HPSG or CCG.

Another interesting reason to tackle this problem is to understand the mechanism of child language acquisition. In particular, since the structural constraint that we impose originally is motivated by psycholinguistic observations (Section 2.2.6), we can regard the current task as controlled experiments to see whether the (memory) limitation that children may suffer from may in turn facilitate the acquisition of language. This is however not our primary motivation since there are large gaps between the actual environment in which children acquire language and the current task; see Section 1.1.2 for the detailed discussion. We therefore think the current study to be a starting point for the modeling of a more realistic acquisition scenario, such as the joint inference of word categories and syntax.

As in the previous chapter, this chapter starts with the conceptual part, in which the main focus is the learning algorithm with structural constraints, then followed by the empirical part that focuses on experiments. Our model is basically the dependency model with valence (Klein and Manning, 2004) that we formalized as a special instance of split bilexical grammar (SBG) in Section 2.3.6. We describe how this model can be encoded in a chart parser that simulates left-corner dependency parsing as presented in the previous chapter, which captures the *center-embeddedness* of a subderivation at each chart entry. Intuitively, with this technique we can bias the model to prefer some syntactic patterns, e.g., that do not contain many center-embedding. We discuss the high level idea and mathematical formalization of this approach in Section 5.1 and then present a new chart parsing algorithm that simulates split bilexical grammars in a left-corner parsing strategy 5.2. We then empirically evaluate whether such structural constraints would help to learn good parameters for the model (Sections 5.3 and 5.4). As in the previous chapter, we study this effect across diverse languages; the total number of treebanks that we use is 30 across 24 languages.

Our main empirical finding is that the constraint on center-embeddedness brings at least the same or superior effects as the closely related structural bias on dependency *length* (Smith and Eisner, 2006), i.e., the preference for *shorter* dependencies. In particular, we find that our bias often outperforms length-based ones when additional syntactic cues are given to the model, such as the one that the sentence root should be a verb or a noun. For example, when such a constraint on the root POS tag is given, our method that penalizes center-embeddedness achieves an attachment score of 62.0 on Google universal treebanks (averaged across 10 languages, evaluated on length ≤ 10 sentences), which is superior to the model with the bias on dependency length (58.6) and the model utilizing a larger number of hand crafted rules between POS tags (56.0) (Naseem et al., 2010).

5.1 Approach Overview

5.1.1 Structure-Constrained Models

Every model presented in this section can be formalized as the following joint model over a sentence x and a parse tree z :

$$p(x, z|\theta) = \frac{p_{\text{ORIG}}(x, z|\theta) \cdot f(z, \theta)}{Z(\theta)} \quad (5.1)$$

where $p_{\text{ORIG}}(x, z|\theta)$ is a (baseline) model, such as DMV. $f(z, \theta)$ assigns a value between $[0, 1]$ for each z , i.e., it works as a penalty or a cost, *reducing* the original probability depending on z . One such penalty that we consider is prohibiting any trees that contain any center-embedding, which is represented as follows:

$$f(z, \theta) = \begin{cases} 1 & \text{if } z \text{ contains no center-embedding;} \\ 0 & \text{otherwise.} \end{cases} \quad (5.2)$$

In Section 5.2, we present a way to encode such a penalty term during the CKY-style algorithm.

Though $f(z, \theta)$ works as adding a penalty to each original probability, the distribution $p(x, z|\theta)$ is still normalized; here $Z(\theta) = \sum_{x, z} p_{\text{ORIG}}(x, z|\theta) \cdot f(z, \theta)$.

Intuitively, $f(z, \theta)$ models the preferences that the original model $p_{\text{ORIG}}(x, z|\theta)$ does not explicitly encode. Note that we do not try to learn $f(z, \theta)$; every constraint is given as an *external* constraint.

Note that this simple approach to combine two models is not entirely new and has been explored several times. For example, Pereira and Schabes (1992) present an EM algorithm that relies on partially bracketed information and Smith and Eisner (2006) model $f(z, \theta)$ as the dependency length-based penalty term. We explore several kinds of $f(z, \theta)$ in our experiments including the existing one, e.g., dependency length, and our new idea, center-embeddedness, examining which kind of structural constraint is most helpful for learning grammars in a cross-linguistic setting. Below, we discuss the issues on learning of this model. The main result was previously shown in Smith (2006) though we summarize it here in our own terms defined in Chapter 2 for completeness.

5.1.2 Learning Structure-Constrained Models

At first glance, the normalization constant $Z(\theta)$ in Eq. 5.1 seems to prevent the use of the EM algorithm for parameter estimation for this model. We show here that in practice we need not care about this constant and the resulting EM algorithm will increase the likelihood of the model of Eq. 5.1.

Recall that the EM algorithm collects expected counts for each rule r , $e(r|\theta)$ at each E-step and then normalizes the counts to update the parameters. We decomposed $e(r|\theta)$ into the counts for each span of each sentence as follows:

$$e(r|\theta) = \sum_{x \in \mathbf{x}} e_x(r|\theta) = \sum_{x \in \mathbf{x}} \sum_{0 \leq i \leq k \leq j \leq n_x} e_x(z_{i,k,j,r}|\theta). \quad (5.3)$$

We now show that correct $e_x(z_{i,k,j,r}|\theta)$ under the model (Eq. 5.1) is obtained without a need to compute $Z(\theta)$. Let $q(x, z|\theta) = p_{\text{ORIG}}(x, z|\theta) \cdot f(z, \theta)$ be an *unnormalized* (i.e., deficient) distribution over x and z . Then $p(x, z|\theta) = q(x, z|\theta)/Z(\theta)$. Note that we can use the inside-outside algorithm to collect counts under the deficient distribution $q(x, z|\theta)$. For example, we can obtain the (deficient) sentence marginal probability $q(x|\theta) = \sum_{z \in \mathcal{Z}(x)} q(x, z|\theta)$ by modifying rule probabilities appropriately. More specifically, in the case of eliminating center-embedding, our chart may record the current stack depth at each chart entry that corresponds to some subderivation, and then assign zero probability to a chart entry if the stack depth exceeds some threshold.

We can represent $e_x(z_{i,k,j,r}|\theta)$ using $q(x, z|\theta)$ instead of $p(x, z|\theta)$, which is more complex. The key observation is that each $e_x(z_{i,k,j,r}|\theta)$ is represented as the ratio of two quantities:

$$e_x(z_{i,k,j,r}|\theta) = \frac{p(z_{i,k,j,r} = 1, x|\theta)}{p(x|\theta)}. \quad (5.4)$$

Calculating these quantities is hard due to the normalization constant. However, as we show below, the normalization constant is canceled in the course of computing the ratio, meaning that the expected counts (under the correct distribution) are obtained with the inside-outside algorithm under the unnormalized distribution $q(x, z|\theta)$. Let us first consider the denominator in Eq. 5.4, which can

be rewritten as follows:

$$p(x|\theta) = \sum_{z \in \mathcal{Z}(x)} p(x, z|\theta) = \sum_{z \in \mathcal{Z}(x)} \frac{q(x, z|\theta)}{Z(\theta)} = \frac{q(x|\theta)}{Z(\theta)}. \quad (5.5)$$

For the numerator, we first observe that

$$p(z_{i,k,j,r} = 1, x|\theta) = \sum_{z \in \mathcal{Z}(x)} p(z_{i,k,j,r} = 1, z, x|\theta) \quad (5.6)$$

$$= \sum_{z \in \mathcal{Z}(x)} p(z, x|\theta) p(z_{i,k,j,r} = 1|z) \quad (5.7)$$

$$= \sum_{z \in \mathcal{Z}(x)} p(z, x|\theta) \mathbb{I}(z_{i,j,k,r} \in z) \quad (5.8)$$

$$= \frac{\sum_{z \in \mathcal{Z}(x)} q(x, z|\theta) \mathbb{I}(z_{i,j,k,r} \in z)}{Z(\theta)}, \quad (5.9)$$

where $\mathbb{I}(c)$ is an identity function that returns 1 if c is satisfied and 0 otherwise. The numerator in Eq. 5.9 is the value that the inside-outside algorithm calculates for each $z_{i,j,k,r}$ (with Eq. 2.17), which we write as $q(z_{i,k,j,r} = 1, x|\theta)$. Thus, we can skip computing the normalization constant in Eq. 5.4 by replacing the quantities with the ones under $q(x, z|\theta)$ as follows:

$$e_x(z_{i,k,j,r}|\theta) = \frac{p(z_{i,k,j,r} = 1, x|\theta)}{p(x|\theta)} = \frac{q(z_{i,k,j,r} = 1, x|\theta)}{q(x|\theta)}. \quad (5.10)$$

The result indicates that by running the inside-outside algorithm *as if* our model is deficient, using $q(x, z|\theta)$ in place of $p(z, x|\theta)$, we can obtain the model with higher likelihood of $p(x, z|\theta)$ (Eq. 5.1). Note that the viterbi parse can also be obtained using $q(x, z|\theta)$ since $\arg \max_z q(x, z|\theta) = \arg \max_z p(x, z|\theta)$ holds.

5.2 Simulating split-bilexical grammars with a left-corner strategy

Here we present the main theoretical result in this chapter. In Section 2.3 we showed that the parameters of the very general model for dependency trees called split-bilexical grammars (SBGs) can be learned using the EM algorithm with CKY-style inside-outside calculation. Also, we formalized the left-corner dependency parsing algorithm as a transition system in Chapter 4, which enables capturing the *center-embeddedness* of the current derivation via *stack depth*. We combine these two parsing techniques in a non-trivial way, and obtain a new chart parsing method for split-bilexical grammars that enables us to calculate center-embeddedness of each subderivation at each chart entry.

We describe the algorithm based on the inference rules with items (as the triangles in Section 2.3.5). The basic idea is that we *memoize* the subderivations of left-corner parsing, which share the same information and look the same under the model. Basically, each chart item is a stack element;

for example, an item $\triangle_{i \ h \ j}$ abstracts (complete) subtrees on the stack headed by h spanning i to j . Each inference rule then roughly corresponds to an action of the transition system.¹ Thus, if we extract one derivation from the chart, which is a set of inference rules, it can be mapped to a particular sequence of transition actions. On this chart, each item is further decorated with the current stack depth, which is the key to capture the center-embeddedness efficiently during dynamic programming.

A particular challenge for efficient tabulation is similar to the one that we discussed in Section 2.3.5; that is, we need to eliminate the spurious ambiguity for correct parameter estimation and for reducing time complexity. In Section 5.2.3, we describe how this can be achieved by applying the idea of head-splitting into the tabulation of left-corner parsing. In the following two sections we discuss some preliminaries for developing the algorithm, i.e., how to handle dummy nodes on the chart (Section 5.2.1) and a note on parameterization of SBGs with the left-corner strategy (Section 5.2.2).

5.2.1 Handling of dummy nodes

An obstacle when designing chart items abstracting many derivations is the existence of predicted nodes, which were previously abstracted with *dummy* nodes in the transition system. Unfortunately, we cannot use the same device in our dynamic programming algorithm because it leads to very inefficient asymptotic runtime. Figure 5.1 explains the reason for this inefficiency. In the transition system, we postponed scoring of attachment preferences between a dummy token and its left dependents (e.g., Figure 5.1(a)) until *filling* the dummy node with an actual token by an INSERT action; this mechanism makes the algorithm fully incremental, though it requires remembering every left dependent token (see INSERT in Figure 4.4) at each step. This tracking of child information is too expensive for our dynamic programming algorithm. To solve this problem, we instead fill a dummy node with an actual token when the dummy is first introduced (not when INSERT is performed). This is impossible in the setting of a transition system since we do not observe the unread tokens in the portion of the sentence remaining in the buffer. Figure 5.1(c) shows an example of an item used in our dynamic programming, which does not abstract the predicted token as a dummy node, but abstracts the construction of child subtrees spanning i to j below the predicted node p . An arc from p indicates that at least one token between i and j is a dependent of p , although the number of dependents as well as the positions are unspecified.

5.2.2 Head-outward and head-inward

The generative process of SBGs described in Section 2.3.5 is *head-outward*, in that its state transition $q_1 \xrightarrow{a} q_2$ is defined as a process of *expanding* the tree by generating a new symbol a which is the most distant from its head when the current state is q_1 . The process is called *head-inward* (e.g., Alshawi (1996)) if it is reversed, i.e., when the closest dependent of a head on each side is generated last. Note that the generation process of left-corner parsing cannot be described fully

¹We also introduce extra rules, which are needed for encoding parameterization of SBGs, or achieving head-splitting as we describe later.

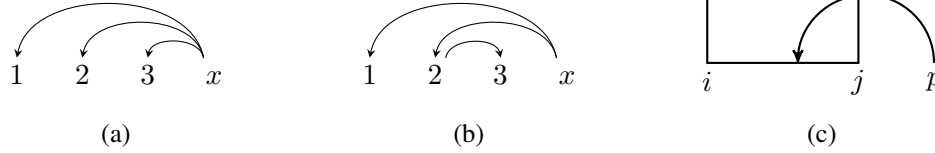


Figure 5.1: Dummy nodes (x in (a) and (b)) in the transition system cannot be used in our transition system because with this method, we have to remember every child token of the dummy node to calculate attachment scores at the point when the dummy is filled with an actual token, which leads to an exponential complexity. We instead abstract trees in a different way as depicted in (c) by not abstracting the predicted node p but filling with the actual word (p points to some index in a sentence such that $j < p \leq n$). If $i = 1, j = 3$, this representation abstracts both tree forms of (a) and (b) with some fixed x (corresponding to p).

head-outward. In particular, its generation of left dependents of a head is inherently head-inward since a parser builds a tree from left to right. For example, the tree of Figure 5.1(a) is constructed by first doing LEFTPRED when 1 is recognized, and then attaching 2 and 3 in order. Fortunately, these two processes, head-inward and head-outward, can generally be interchanged by reversing transitions (Eisner and Satta, 1999). In the algorithm described below, we model its left automaton L_a as a head-inward process while right automaton R_a as a head-outward process. Specifically, that means if we write $q_1 \xrightarrow{a'} q_2 \in L_a$, the associated weight for this transition is $p(a'|q_2)$ instead of $p(a'|q_1)$. We also do not modify the meaning of sets $final(L_a)$ and $init(L_a)$; i.e., the left state is initialized with $q \in final(L_a)$ and finishes with $q \in init(L_a)$.

5.2.3 Algorithm

Figures 5.2 and 5.3 describe the algorithm that parses SBGs with the left-corner parsing strategy. Each inference rule can be basically mapped to a particular action of the transition, though the mapping is sometimes not one-to-one. For example SHIFT action is divided into two cases, LEFT and RIGHT, for achieving head-splitting. Some actions, e.g., INSERT (INSERT-LEFT and INSERT-RIGHT) and LEFTCOMP (LEFTCOMP-L-* and LEFTCOMP-R-*) are further divided into two cases depending on tree structure of a stack element, i.e., whether a predicted node (a dummy) is a head of the stack element or some right dependent.

Each chart item preserves the current stack depth d . The algorithm only accepts an item spanning the whole sentence (including the dummy root symbol $\$$ at the end of the sentence) with the stack depth one and this condition certifies that the derivation can be converted to a valid sequence of transitions. When our interest is to eliminate derivations that contain the specific depth of center-embedding, it can be achieved by assigning zero weight to every chart cell in which the depth exceeds the threshold.

See LEFTCOMP-L-1 and LEFTCOMP-L-2 (LEFTCOMP-R-* is described later); these are the points where deeper stack depth might be detected. These two rules are the result of decomposing a single LEFTCOMP action in the transition system into the *left phase*, which collects only left half constituent given a head h , and *right phase*, which collects the remaining. Figure 5.4 describes how

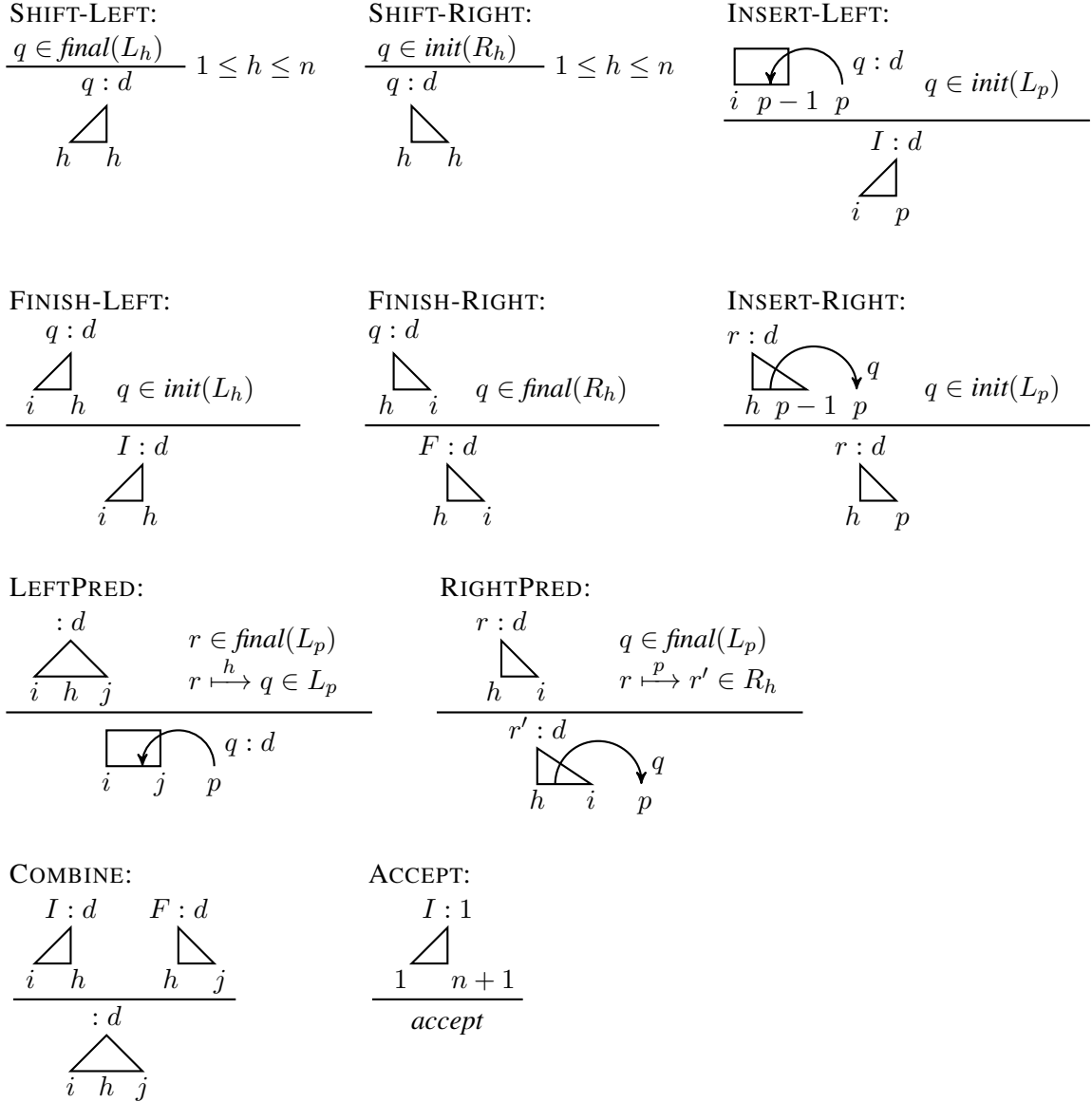


Figure 5.2: An algorithm for parsing SBGs with a left-corner strategy in $O(n^4)$ given a sentence of length n , except the composition rules which are summarized in Figure 5.3. The $n + 1$ -th token is a dummy root token $\$$, which only has one left dependent (sentence root). i, j, h, p are indices of tokens. The index of a head which is still collecting its dependents is decorated with a state (e.g., q). L_h and R_h are left and right FSAs of SBGs given a head index h , respectively; we reverse the process of L_h to start with $q \in \text{final}(L_h)$ and finish with $q \in \text{init}(L_h)$ (see the body). Each item is also decorated with depth d that corresponds to the stack depth incurred when building the corresponding tree with left-corner parsing. Since an item with larger depth is only required for composition rules, the depth is unchanged with the rules above, except SHIFT-*, which corresponds to SHIFT transition and can be instantiated with arbitrary depth. Note that ACCEPT is only applicable for an item with depth 1, which guarantees that the successful parsing process remains a single tree on the stack. Each item as well as a statement about a state (e.g., $r \in \text{final}(L_p)$) has a weight and the weight of a consequence item is obtained by the product of the weights of its antecedent items.

LEFTCOMP-L-1:

$$\frac{
 \begin{array}{c}
 \boxed{i \quad j} \xrightarrow{q:d} p \quad \triangle_{j+1 \quad h}^{I:d} \\
 \hline
 b
 \end{array}
 }{
 \boxed{i \quad h} \xrightarrow{q:d} p
 }
 \quad b = \begin{cases} 1 & \text{if } h - (j+1) \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

LEFTCOMP-L-2:

$$\frac{
 \begin{array}{c}
 \boxed{i \quad h} \xrightarrow{q:d} p \quad \triangle_{h \quad j}^{F:d'} \quad q \xrightarrow{h} q' \in L_p \\
 \hline
 b
 \end{array}
 }{
 \boxed{i \quad j} \xrightarrow{q':d} p
 }
 \quad d' = \begin{cases} d+1 & \text{if } b=1 \text{ or } (j-h) \geq 1 \\ d & \text{otherwise.} \end{cases}$$

LEFTCOMP-R-1:

$$\frac{
 \begin{array}{c}
 \triangle_{i \quad j}^{r:d} \xrightarrow{q} p \quad \triangle_{j+1 \quad h}^{I:d} \\
 \hline
 b
 \end{array}
 }{
 \triangle_{i \quad h}^{r:d} \xrightarrow{q} p
 }
 \quad b = \begin{cases} 1 & \text{if } h - (j+1) \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

LEFTCOMP-R-2:

$$\frac{
 \begin{array}{c}
 \triangle_{i \quad h}^{r:d} \xrightarrow{q} p \quad \triangle_{h \quad j}^{F:d'} \quad q \xrightarrow{h} q' \in L_p \\
 \hline
 b
 \end{array}
 }{
 \triangle_{i \quad j}^{r:d} \xrightarrow{q'} p
 }
 \quad d' = \begin{cases} d+1 & \text{if } b=1 \text{ or } (j-h) \geq 1 \\ d & \text{otherwise.} \end{cases}$$

RIGHTCOMP:

$$\frac{
 \begin{array}{c}
 \triangle_{i \quad h-1 \quad h}^{r:d} \xrightarrow{q} p \quad \triangle_{h \quad j}^{q':d'} \\
 \hline
 b
 \end{array}
 }{
 \triangle_{i \quad j}^{r:d} \xrightarrow{s} p
 }
 \quad \begin{array}{l} q \in \text{init}(L_h) \\ q' \xrightarrow{p} q'' \in \text{final}(R_h) \\ s \in \text{final}(L_p) \end{array}
 \quad d' = \begin{cases} d+1 & \text{if } (j-h) \geq 1 \\ d & \text{otherwise.} \end{cases}$$

Figure 5.3: The composition rules that are not listed in Figure 5.2. LEFTCOMP is divided into two cases, LEFTCOMP-L-* and LEFTCOMP-R-* depending on the position of the dummy (predicted) node on the left antecedent item (corresponding to the second top element on the stack). They are further divided into two processes, 1 and 2 for achieving head-splitting. b is an additional annotation on an intermediate item for correct depth computation in LEFTCOMP.

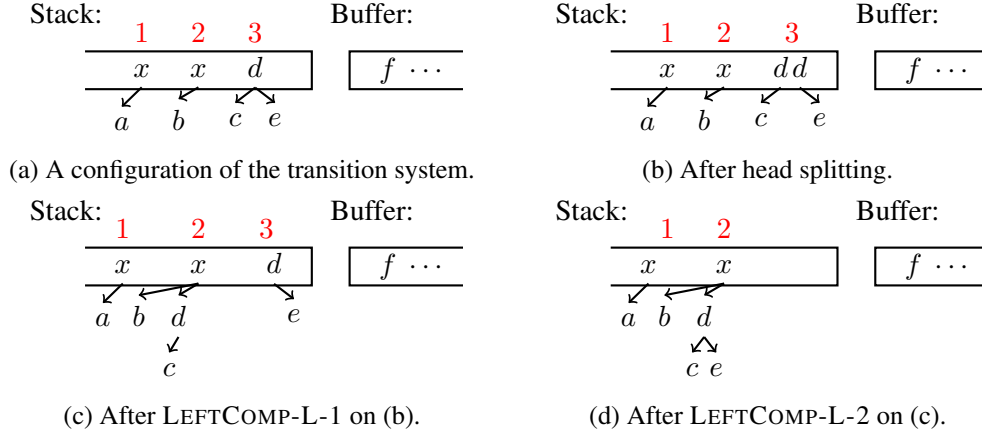


Figure 5.4: We decompose the LEFTCOMP action defined for the transition system into two phases, LEFTCOMP-L-1 and LEFTCOMP-L-2, each of which collects only left or right half constituent of a subtree on the top of the stack. A number above each stack element is the stack depth decorated on the corresponding chart item.

this decomposition looks like in the transition system. As shown in Figure 5.4(b), we imagine that the top subtree on the stack is divided into a left and right constituents.² In Figure 5.4 we number each subtree on the stack from left to right. Note that then the number of the rightmost (top) element on the stack corresponds to the stack depth of the configuration. This value corresponds to the depth


$$F : d'$$

annotated on each item in the algorithm, such as d' in \triangleleft appeared as the right antecedent item of LEFTCOMP-L-2 in Figure 5.3. Then, since the left antecedent item of the same rule, i.e., $\square \curvearrowright$, corresponds to the second top element on the stack, its depth d is generally smaller by one, i.e., $d = d' - 1$.





One complicated point with this depth calculation is that larger stack depth should not always be detected during this computation. Recall the discussion in Section 4.3.4 that there are two kinds of stack depth that we called depth_{re} and depth_{sh} , in which only depth_{re} correctly captures the center-embeddedness of a construction. Depth_{sh} is the depth of a configuration after a shift action, on which the top element is *complete*, i.e., contains no dummy (predicted) node. Note that the depth annotated on each subtree in Figure 5.3 is in fact depth_{sh} , as the right antecedent item of each rule (corresponds to the top element on the stack) does not contain a predicted node. Since our goal is to capture the center-embeddedness during parsing, we fix this discrepancy with a small trick, which is described as the side condition of LEFTCOMP-L-1 and LEFTCOMP-L-2 in Figure 5.3. The point at which only depth_{sh} increases by 1 is when a shifted token is immediately reduced with a following composition rule. We treat this process as a special case and do not increase the stack depth when the span length of the subtree that is reduced with a composition rule (right antecedent) is 1.

² This assumption does not change the incurred stack depth at each step since in the left-corner transition system, right half dependents of a head are collected only after its left half construction is finished. Splitting a head as in Figure 5.4(b) means that we treat these left and right parsing processes independently.

The remained problem is that because we split each constituent into left and right constituents, we cannot calculate the size of the reduced constituent immediately. The additional variable b in Figure 5.3 is introduced for checking this condition. b is set to 1 if the left part, i.e., LEFTCOMP-L-1 collects a constituent with span length greater than one ($h = j + 1$ indicates one word constituent). If $b = 1$, regardless of the size of remaining right constituent, the second phase, or the right part LEFTCOMP-L-2 always increases the stack depth (i.e., $d' = d + 1$). $b = 0$ means the size of left constituent is zero; in this case, the stack depth is increased when the size of the right constituent is greater than one. In summary, the stack depth of the right antecedent item is increased *unless* three indices, $j + 1$ and h in LEFTCOMP-L-1 and j in LEFTCOMP-L-2 are identical, which occurs when the reduced constituent is a single shifted token.

Finally, we note that we do not modify the depth of the right antecedent item of LEFTCOMP-L-1. This is correct since the consequent item of this rule , which waits for a right half constituent of h , can only be used as an antecedent of the following LEFTCOMP-L-2 rule. The role of LEFTCOMP-L-1 is just to annotate the head of the reduced constituent and the span length. Then LEFTCOMP-L-2 checks the depth condition and calculates the weight associated with LEFTCOMP action (i.e., $q \xrightarrow{h} q' \in L_p$).


The other part of the algorithm can be understood as follows:

- LEFTCOMP-R-* is almost the same as LEFTCOMP-L-*. The difference is in the tree shape of the left antecedent item; in the R case, it is a right half constituent with a predicted node , which corresponds to a subtree in which the predicted node is not the head but the tail of the right spine. We distinguish these two trees in the algorithm since they often behave very differently as shown in Figure 5.2. Note that  has two FSA states since it contains two head tokens collecting its dependents (i.e., the head of the tree and the predicted token).
- Differently from LEFTCOMP, RIGHTCOMP is summarized as a single rule while it seems a bit complicated. In the algorithm, RIGHTCOMP can only be performed when the predicted token of  finishes collecting its left dependents (indicated as the consecutive indices of $h - 1$ and h). See Section 5.2.4 for the reason of this restriction. Another condition for applying this rule is that the right state q' of head h must be a final state after applying transition $q' \xrightarrow{p} q''$, which collects new rightmost dependent of h , i.e., p . Under these conditions, the rule performs the following parser actions: 1) attach p as a right dependent of h ; 2) finishes the right FSA of h ; and 3) start collecting left dependents of p by setting the final state to the left state of it.
- Some rules such as FINISH-* and COMBINE do not exist in the original transition system. We introduce these to represent the generative process of SBGs in the left-corner algorithm.
- We do not annotate a state on a triangle . This is because it can only be deduced by COMBINE, which combines two finished constituents with the same head.
- A parse always finishes with a consecutive application of LEFTPRED, INSERT-LEFT, and ACCEPT after a parse spanning the sentence $1 \triangle_n$ is recognized. LEFTPRED predicts

the arc between the dummy root token \$ at $n + 1$ -th position and this parse tree, and then INSERT-LEFT removes this predicted arc. Note that to get a correct parse the left FSA for \$ should be modified appropriately for not collecting more than one dependent (the common parameterization of DMV automatically achieve this).

5.2.4 Spurious ambiguity and stack depth

We have splitted each head token into left and right, which means each derivation with this algorithm has one-to-one correspondence to a dependency tree. That is, there is no spurious ambiguity and the EM algorithm works correctly (Section 2.3.4). In Section 4.3.3, we developed an oracle for a transition system that returns a sequence of gold actions given a sentence and a dependency tree, and found that the presented oracle is *optimal* in terms of incurred stack depth. This optimality essentially comes from the implicit binarization mechanism of the oracle given a dependency tree (Theorem 4.1).

The chart algorithm presented above has the same mechanism of binarization, and thus is optimal in terms of incurred stack depth. Essentially this is due to our design of LEFTCOMP and RIGHTCOMP in Figure 5.3. We do not allow RIGHTCOMP for a rectangle  as in the case of LEFTCOMP-L-*. Also, there is no second phase in RIGHTCOMP, meaning that the reduced constituent (i.e., the top stack element) does not have left children. We notice that these two conditions are exactly the same as the statement in Lemma 4.1, which was the key to prove the binarization mechanism in Theorem 4.1. When we are interested in another *nonoptimal* parsing algorithm, what we should do is to modify the allowed condition for LEFTCOMP and RIGHTCOMP, e.g., perhaps RIGHTCOMP is divided into several parts and instead the condition that LEFTCOMP can be applied is highly restricted.

5.2.5 Relaxing the definition of center-embedding

In the previous chapter, we have examined a simple relaxation for the definition of center-embedding by allowing constituents up to some length to be at the top on the stack. Here we demonstrate how this relaxation can be implemented with a simple modification to the presented algorithm.

Let us assume the situation in which we allow constructions with one degree of center-embedding *if* the length of embedded constituent is at most three; that is, we allow a small part of center-embedding. We write this as $(D, C) = (1, 3)$, meaning that the maximum stack depth is generally one ($D = 1$) though we partially allow depth two when the length of the embedded constituent is less than or equal to three ($C = 3$). This can be achieved by modifying the role of variable b and equations in Figure 5.3. As we have seen, the current algorithm does not increase the stack depth of right antecedent item with LEFTCOMP or RIGHTCOMP if the length of those reduced constituents is just one (has no dependent), which corresponds to the case of $C = 1$. Our goal is to generalize this calculation and to judge whether the length of the reduced constituent is greater than C or not. We modify the side condition of LEFTCOMP-* -1 as follows:

$$b = \max(C, h - (j + 1)). \quad (5.11)$$

Now b is a variable in the range $[0, C]$. $b = 0$ means the left constituent is one word. Then, the side

condition of LEFTCOMP-2 is changed as follows:

$$d' = \begin{cases} d + 1 & \text{if } b + (j - h) \geq C \\ d & \text{otherwise.} \end{cases} \quad (5.12)$$

For example, when $C = 3$, and the left constituent is ${}_3\triangleleft_4$ while the right constituent is ${}_4\triangleleft_5$, the depth is unchanged since $b + (j - h) = 2 < 3$ in Eq. 5.12. Note that the algorithm may be inefficient when C is larger, although it is not a practical problem as we only explore very small values such as 2 and 3.

5.3 Experimental Setup

Now we move on to the empirical part of this chapter. This section summarizes the experimental settings such as the datasets that we use, evaluation method, and possible constraints that we impose to the models. In particular, we point out the crucial issue in the current evaluation metric in Section 5.3.3 and then propose our solution to alleviate this problem in Section 5.3.4.

5.3.1 Datasets

We use two different multilingual corpora for our experiments: Universal Dependencies (UD) and Google universal dependency treebanks; the characteristics of these two corpora are summarized in Chapter 3. We mainly use UD in this chapter, which comprises of 20 different treebanks. One problem of UD is that because this is the first study (in our knowledge) to use it in unsupervised dependency grammar induction, we cannot compare our models to previous state-of-the-art approaches. The Google treebanks are used for this purpose. It comprises of 10 languages and we discuss the relative performance of our approaches compared to the previously reported results on this dataset.

Preprocess Some treebanks of UD are annotated with multiword expressions although we strip them off for simplicity. Also we remove every punctuation mark in every treebank (both training and testing). This preprocessing has been performed in many previous studies. This is easy when the punctuation is at a leaf position; otherwise, we reattach every child token of that punctuation to its closest ancestor that is not a punctuation.

Input token Every model in this chapter only receives annotated part-of-speech (POS) tags given in the treebank. This is a crucial limitation of the current study both from the practical and cognitive points of view as we discussed in Section 2.4. We learn the model on the *unified*, universal tags given in the respective corpora. In Google treebanks, the total number of tags is 11 (excluding punctuation) while that is 16 in UD. See Chapter 3 for more details.

Sentence Length Often unsupervised parsing systems are trained and tested on a subset of the original training/testing set by setting a maximum sentence length and ignoring every sentence

| Language | #Sents. | #Tokens | Av. len. | Test ratio |
|------------|---------|---------|----------|------------|
| Basque | 3,743 | 31,061 | 8.2 | 24.9 |
| Bulgarian | 6,442 | 53,737 | 8.3 | 11.0 |
| Croatian | 1,439 | 15,285 | 10.6 | 6.6 |
| Czech | 46,384 | 388,309 | 8.3 | 12.9 |
| Danish | 2,952 | 25,455 | 8.6 | 6.5 |
| English | 9,279 | 67,249 | 7.2 | 14.9 |
| Finnish | 10,146 | 85,057 | 8.3 | 5.2 |
| French | 5,174 | 55,413 | 10.7 | 2.1 |
| German | 8,073 | 82,789 | 10.2 | 6.7 |
| Greek | 746 | 6,987 | 9.3 | 11.2 |
| Hebrew | 1,883 | 19,057 | 10.1 | 9.7 |
| Hungarian | 580 | 5,785 | 9.9 | 11.0 |
| Indonesian | 2,492 | 25,731 | 10.3 | 11.5 |
| Irish | 408 | 3,430 | 8.4 | 18.6 |
| Italian | 5,701 | 51,272 | 8.9 | 4.3 |
| Japanese | 2,705 | 28,877 | 10.6 | 25.9 |
| Persian | 1,972 | 18,443 | 9.3 | 9.9 |
| Spanish | 4,249 | 45,608 | 10.7 | 2.2 |
| Swedish | 3,545 | 31,682 | 8.9 | 20.7 |

Table 5.1: Statistics on UD15 (after stripping off punctuations). Av. len. is the average length. Test ratio is the token ratio of the test set.

longer than the threshold. The main reason for this filtering during training is efficiency: running dynamic programming ($O(n^4)$ in our case) for longer sentences in many number of iterations is expensive. We therefore set the maximum sentence length during training to 15 on UD experiments, which is not so expensive to explore many parameter settings and languages. We also evaluate our models against test sentences up to length 15. We choose this value because we are interested more in whether our structural constraint helps to learn basic word order of each language, which may be obscured if we use full sentences as in the supervised parsing experiments since longer sentences are typically conjoined with several clauses. This setting has been previously used in, e.g., Bisk and Hockenmaier (2013). We call this filtered dataset UD15 in the following.

We use different filtering for Google treebanks and set the maximum length for training and testing to 10. This is the setting of Grave and Elhadad (2015), which compares several models including the previous state-of-the-art method of Naseem et al. (2010). See Tables 5.1 and 5.2 for the statistics of the datasets.

5.3.2 Baseline model

Our baseline model is the featurized DMV model (Berg-Kirkpatrick et al., 2010), which we briefly described in Section 2.3.7. We choose this model as our baseline since it is very simple yet performs competitively to the more complicated state-of-the-art systems. Other more sophisticated

| Language | #Sents. | #Tokens | Av. len. | Test ratio |
|---------------|---------|---------|----------|------------|
| German | 3,036 | 23,833 | 7.8 | 8.9 |
| English | 4,341 | 31,287 | 7.2 | 5.7 |
| Spanish | 1,258 | 9,731 | 7.7 | 3.2 |
| French | 1,629 | 13,221 | 8.1 | 2.7 |
| Indonesian | 799 | 6,178 | 7.7 | 10.7 |
| Italian | 1,215 | 9,842 | 8.1 | 6.1 |
| Japanese | 5,434 | 32,643 | 6.0 | 3.6 |
| Korean | 3,416 | 21,020 | 6.1 | 7.7 |
| Br-Portuguese | 1,186 | 9,199 | 7.7 | 11.7 |
| Swedish | 1,912 | 12,531 | 6.5 | 18.5 |

Table 5.2: Statistics on Google trebanks (maximum length = 10).

methods exist, but they typically require much complex inference techniques (Spitkovsky et al., 2013) or external information (Mareček and Straka, 2013), which obscure the contribution of our imposing constraints. Studying the effect of the structural constraints for these more strong models is remained for the future work.

This model contains two tunable parameters, the regularization parameter and the feature templates. We fix the regularization parameter to 10, which is the same as the value in Berg-Kirkpatrick et al. (2010) since we did not find significant performance changes with this value in our preliminary study. We also basically use the same feature templates as Berg-Kirkpatrick et al.; the only difference is that we add additional backoff features for STOP probabilities that ignore both direction and adjacency, which we found slightly improves the performance.

5.3.3 Evaluation

Evaluation is one of the *unsolved* problems in the unsupervised grammar induction task. The main source of difficulty is the inherent ambiguity of the notion of *heads* in dependency grammar that we mentioned several times in this thesis (see Section 3.1 for details). Typically the quality of the model is evaluated in the same way as the supervised parsing experiments: At test time, the model predicts dependency trees on test sentences; then the *accuracy* of the prediction is measured by an *unlabelled attachment score* (UAS):

$$\text{UAS} = \frac{\# \text{ tokens whose head matches the gold head}}{\# \text{ tokens}} \quad (5.13)$$

The problem of this measure is that it completely ignores the ambiguity of head definitions since its score calculation is against the single gold dependency treebank. Some attempts to alleviate the problem of UAS exist, such as direction-free (undirected) measure (Klein and Manning, 2004) and a more sophisticated measure called neutral edge detection (NED) (Schwartz et al., 2011). NED expands the set of *correct* dependency constructions given the predicted tree and the gold tree to save the errors that seem to be caused by annotation variations. However NED is a too lenient metric and causes different problems. For example, under NED (also under the undirected measure) the

two trees $\text{dogs} \hat{\curvearrowright} \text{ran}$ and $\text{dogs} \hat{\curvearrowright} \text{ran}$ are treated equal, although it is apparent that the correct analysis is the former. We suspect this is the reason why many researchers have not used NED and instead select UAS while recognizing the inherent problems (Cohen, 2011; Bisk and Hockenmaier, 2013).

However, the current situation is really unhealthy for our community. For example, if we find some method that can boost UAS from 40 to 60, we cannot identify whether this improvement is due to the acquisition of essential word orders such as dependencies between nouns and adjectives, or just overfitting to the current gold treebank. The latter case occurs, e.g., when the current gold treebank assumes that heads of prepositional phrases are the content words and the improved model changes the analysis for them from functional heads to content heads. Since our goal is not to obtain a model that can overfit to the gold treebank in a surface level, but to understand the mechanism that the model can acquire better word orders, we want to remove the possibility to make such (fake) improvements in our experiments.

We try to minimize this problem not by revising the evaluation metric but by customizing models. We basically use UAS since the other metrics have more serious drawbacks. However, to avoid unexpected improvements/degradations, we constraint the model to explore only trees that may follow the conventions in the current gold data. This is possible in our corpora as they are annotated under some consistent annotation standard (see Chapter 3). This approach is conceptually similar to Naseem et al. (2010), although we do not incorporate many constraints on word orders, such as the dependencies between a verb and a noun. The detail of the constraints we impose to the models is described next.

5.3.4 Parameter-based Constraints

The goal of the current experiments is to see the effect of *structural constraint*, which we hope to guide the model to find better parameters. To do so, on the baseline model (Section 5.3.2) we impose several additional constraints in the framework of structural constraint model described in Section 5.1, and examine how performance changes (we list these constraints in Section 5.3.5).

In addition to the structural constraints, we also consider another kind of constraint that we call *parameter-based constraint* in the same framework, that is, as a cost function $f(z, \theta)$ in Eq. 5.2. The parameter-based constraints are constraints on POS tags in a given sentence and are specified declaratively, e.g., *X cannot have a dependent in the sentence*. Note that our main focus in this experiment is the effect of structural constraints. As we describe below, the parameter-based ones are introduced to make the empirical comparison of structural constraints more meaningful.

Note that all constraints below are imposed during training only, as we found in our preliminary experiments that the constraints during decoding (at test time) make little performance changes. This is natural in particular for parameter-based constraints since the model learns the parameters that follow the given constraints during training.

We consider the following constraints in this category:

Function words in UD This constraint is introduced to alleviate the problem of evaluation that we discussed in Section 5.3.3. One characteristic of UD is that its annotation style is consistently content head based, that is, every function word is analyzed as a dependent of the most

closely related content word.³ By forcing the model to explore only structures that follow this convention, we expect we can minimize the problem of *arbitrariness* of head choices. This constraint can easily be implemented by setting every STOP probability of DMV for function words to 1. We regard the following six POS tags as function words: ADP, AUX, CONJ, DET, PART, and SCONJ. Since most arbitrary constructions are around function words, we hope this makes the performance change due to other factors such as the structural constraints clearer. Note that this technique is still not the perfect and cannot neutralize some annotation variations such as internal structures of noun phrases; we do not consider further constraints to save such more complex cases.

Function words in Google treebanks We consider the similar constraints on Google treebanks. The Google treebanks uses the following four POS tags for function words: DET, CONJ, PRT, and ADP. PRT is a particle corresponding to PART in UD. As in UD it also follows the annotation standard of Stanford typed dependencies (McDonald et al., 2013) and analyzes most function words as dependents, although it is not the case for ADP, which is in most cases analyzed as a head of the governing phrase. We therefore introduce another kind of constraint for ADP, which prohibits to become a dependent, i.e., ADP must have at least one dependent. Implementing this constraint in our dynamic programming algorithm is a bit involved compared to the previous *unheadable* constraints, mainly due to our *split-head* representation. We can achieve this constraint in a similar way to the constituent length memoization technique that we introduced in Figure 5.3 with variable b . Specifically, at LEFTCOMP-L-1, we remember whether the reduced head h has at least one dependent if h is ADP; then at LEFTCOMP-L-2, we disallow the rule application if that ADP is recognized as having no dependent. We also disallow COMBINE if the head is ADP and the sizes of two constituents are both 1 (i.e., $i = h = j$). The resulting full constituent would be reduced by LEFTPRED to be some dependent, which is although prohibited. Other function words are in most cases analyzed as dependents so we use the same restriction as the function words in UD.

Candidates for root words This constraint is also parameter based though should be distinguished from the above two. Note that the constraints discussed so far are only for alleviating the problem of annotation variations in that they give no hint for acquiring basic word orders for the model such as the preference of an arc from a verb to a noun. This constraint is intended to give such a hint to the model by restricting possible root positions in the sentence. We consider two types of such constraints. The first one is called the *verb-or-noun constraint*, which restricts the possible root word of the sentence to a verb, or a noun. The second one is called the *verb-otherwise-noun constraint*, which more eagerly restricts the search space by only allowing a verb to become a root, if at least one verb exists in the sentence; otherwise, nouns become a candidate. If both do not exist, then every word becomes a candidate. In both corpora, VERB is the only POS tag for representing verbs. We regard three POS tags, NOUN, PRON, and PROPEN in UD as nouns. In Google treebanks, NOUN and PRON are considered as nouns. This type of knowledge is often employed in the previous unsupervised parsing

³We find small exceptions in each treebank probably due to the remaining annotation errors though they are negligibly small.

models in different ways (Gimpel and Smith, 2012; Gormley and Eisner, 2013; Bisk and Hockenmaier, 2012; Bisk and Hockenmaier, 2013) as *seed knowledge* given to the model. We will see how this simple hint to the target grammar affects the performance.

5.3.5 Structural Constraints

These are the constraints that we focus on in the experiments.

Maximum stack depth This constraint removes parses involving center-embedding up to a specific degree and can be implemented by setting the maximum stack depth in the algorithm in Figures 5.2 and 5.3. We also investigate the relaxation of the constraint with a small constituent that we described in Section 5.2.5. Studying the effect of this constraint is the main topic in the our experiments.

Dependency length bias We also explore another structural constraint that biases the model to prefer shorter dependency length, which has previously been examined in Smith and Eisner (2006). With this constraint, each attachment probability is changed as follows:

$$\theta'_A(a|h, dir) = \theta_A(a|h, dir) \cdot \exp(-\beta_{len} \cdot (|h - d| - 1)), \quad (5.14)$$

where $\theta_{A,h,d}$ is the original DMV parameter attaching d as a dependent of h . Differently from Smith and Eisner (2006), we subtract 1 in each length cost calculation to add no penalty for an arc between adjacent words. As Smith and Eisner (2006) noted, this constraint leads to the following form of $f(z, \theta)$ in Eq. 5.2:

$$f_{len}(z, \theta) = \exp \left(-\beta_{len} \cdot \left(\sum_{1 \leq d \leq n} (|h_{z,d} - d|) - n \right) \right), \quad (5.15)$$

where $h_{z,d}$ is the analyzed head position for a dependent at d . Notice that $\sum_{1 \leq d \leq n} (|h_{z,d} - d|)$ is the sum of dependency lengths in the sentence, which means that this model tries to *minimize* the sum of dependency length in the tree and is closely related to the theory of dependency length minimization, a typological hypothesis that grammars may universally favor shorter dependency length (Gildea and Temperley, 2007; Gildea and Temperley, 2010; Futrell et al., 2015; Gulordava et al., 2015).

Notice that typically center-embedded constructions are accompanied by longer dependencies. However the opposite is generally not the case; there are many constructions that do accompany longer dependencies though do not contain center-embedding. By comparing these two constraints, we discuss whether center-embedding is the phenomena worth considering than the simpler method of shorter length bias.

5.3.6 Other Settings

Initialization Much previous works of unsupervised dependency induction, in particular DMV and related models, relied on heuristic initialization called *harmonic initialization* (Klein and Manning, 2004; Berg-Kirkpatrick et al., 2010; Cohen and Smith, 2009; Blunsom and Cohn, 2010), which is obtained by running the first E-step of the training by setting every attachment probability between i and j to $(|i - j|)^{-1}$.⁴ Note that this method also biases the model to favor shorter dependencies.

We do not use this initialization with our structural constraints since one of our motivation is to invent a method that does not rely on such heuristics highly connected to a specific model (like DMV). We therefore initialize the model to be a uniform model. However, we also compare such uniform + structural constrained models to the harmonic initialized model *without* structural constraints to see the relative strength of our approach.

Decoding As noted above, every constraint introduced so far is only imposed during training. At decoding (test time), we do not consider the bias term of Eq. 5.1 and just run the Viterbi algorithm to get the best parse under the original DMV model.

5.4 Empirical Analysis

We first check the performance differences of several settings on UD and then move on to Google treebanks to compare our approach to the state-of-the-art methods.

5.4.1 Universal Dependencies

When no help is given to the root word We first see how the performance changes when using different length biases or stack depth biases (Section 5.3.5). The parameter-based constraint is only the function word constraint of UD, that is, any function word cannot be a head of others. Figure 5.5 summarizes the results. Although the variance is large, we can make the following observations:

- Often small (weak) length biases (e.g., $\beta_{len} = 0.1$) work better than more strong biases. In many languages, e.g., English, Indonesian, and Croatian, the performance improves with a small bias and degrades as the bias is sharpened. Note that the left most point is the score with no constraints, e.g., just the uniformly initialized model.
- We try five different stack depth bound between depth one and depth two. The result shows in many cases the middle, e.g., 1(2), 1(3), and 1(4) works better. The performance of depth two is almost the same as the no-constraint baseline, meaning that stack depth two is too lenient to restrict the search space during learning. This observation is consistent with the empirical stack depth analysis in the previous chapter (Section 4.4.5).

⁴There is another variant of harmonic initialization (Smith and Eisner, 2006) though we do not explore this since the method described here is the one that is employed in Berg-Kirkpatrick et al. (2010) (p.c.), which is our baseline model (Section 5.3.2).

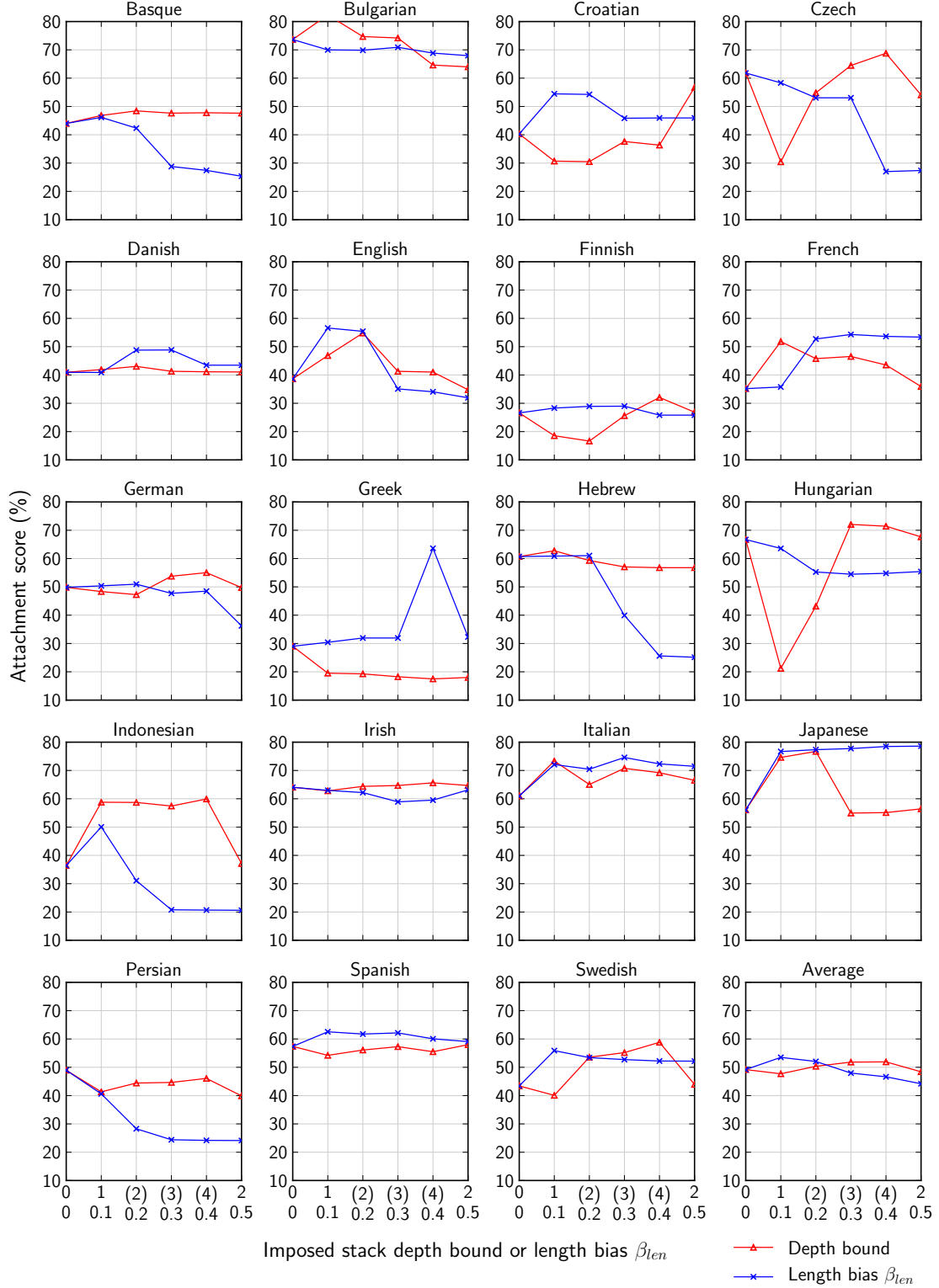


Figure 5.5: Attachment accuracies on UD15 with the function word constraint and structural constraints. The numbers in parentheses are the maximum length of a constituent allowed to be embedded. For example (3) means a part of center-embedding of depth two, in which the length of embedded constituent ≤ 3 , is allowed.

| | Unif. | $C = 2$ | $C = 3$ | $\beta_{len} = 0.1$ | Harmonic |
|------------|-------------|-------------|-------------|---------------------|-------------|
| Basque | 44.0 | 48.5 | 47.6 | 46.1 | 44.5 |
| Bulgarian | 73.6 | 74.7 | 74.2 | 70.0 | 72.5 |
| Croatian | 40.3 | 30.5 | 37.7 | 54.5 | 47.3 |
| Czech | 61.8 | 54.8 | 64.5 | 58.3 | 54.2 |
| Danish | 40.9 | 43.0 | 41.3 | 40.9 | 40.9 |
| English | 38.7 | 54.8 | 41.3 | 56.6 | 39.1 |
| Finnish | 26.6 | 16.7 | 25.6 | 28.3 | 26.4 |
| French | 35.2 | 45.9 | 46.6 | 35.7 | 34.6 |
| German | 49.8 | 47.3 | 53.6 | 50.3 | 49.5 |
| Greek | 29.2 | 19.3 | 18.3 | 30.4 | 49.3 |
| Hebrew | 60.7 | 59.3 | 57.1 | 60.9 | 57.3 |
| Hungarian | 66.7 | 43.2 | 72.2 | 63.6 | 65.8 |
| Indonesian | 36.4 | 58.7 | 57.4 | 50.0 | 40.8 |
| Irish | 64.1 | 64.5 | 64.8 | 63.0 | 64.4 |
| Italian | 61.0 | 65.0 | 70.7 | 72.2 | 65.8 |
| Japanese | 56.1 | 76.7 | 55.0 | 76.7 | 48.2 |
| Persian | 49.0 | 44.4 | 44.7 | 40.7 | 41.4 |
| Spanish | 57.3 | 56.0 | 57.3 | 62.5 | 55.4 |
| Swedish | 43.4 | 53.6 | 55.2 | 55.9 | 49.5 |
| Avg | 49.2 | 50.4 | 51.8 | 53.5 | 49.8 |

Table 5.3: Accuracy comparison on UD15 for selected configurations including harmonic initialization (Harmonic). Unif. is a baseline model without structural constraints. C is the allowed constituent length when the maximum stack depth is one. β_{len} is strength of the length bias.

- On average, we find the best setting is the small length bias $\beta_{len} = 0.1$. In Table 5.3 we summarizes the accuracies of selected configurations in this figure, which work better, as well as the harmonic initialized models.
- The performance for some languages, in particular Greek and Finnish, is quite low compared to other languages. We inspect the output trees for these languages, and found that the model fails to identify very basic word orders, such as the tendency of a verb to be a root word. Essentially, the models so far do not receive such explicit knowledge about grammar, which is known to be particularly hard. We thus see next how performances change if a small amount of *seed knowledge* about the grammar is given the model.

Constraining POS tags for root words Table 5.4 shows the results when we add two kinds of seed knowledge as parameter-based constraints (Section 5.3.4).

We first see the result with the verb-or-noun constraint. This constraint comes from the main assumption of UD that the root token of a sentence is its main predicate, which is basically a verb,

| | Verb-or-noun constraint | | | | Verb-otherwise-noun constraint | | | |
|------------|-------------------------|-------------|-------------|---------------------|--------------------------------|-------------|-------------|---------------------|
| | Unif. | $C = 2$ | $C = 3$ | $\beta_{len} = 0.1$ | Unif. | $C = 2$ | $C = 3$ | $\beta_{len} = 0.1$ |
| Basque | 44.7 | 55.2 | 54.3 | 46.4 | 55.8 | 55.6 | 54.8 | 51.0 |
| Bulgarian | 73.4 | 75.8 | 75.1 | 64.1 | 72.7 | 75.8 | 75.2 | 70.6 |
| Croatian | 40.1 | 52.5 | 41.4 | 47.3 | 57.0 | 52.5 | 52.5 | 55.8 |
| Czech | 50.7 | 54.8 | 64.7 | 59.2 | 63.2 | 54.9 | 66.3 | 58.1 |
| Danish | 40.9 | 43.1 | 41.3 | 40.9 | 48.7 | 46.9 | 50.1 | 47.3 |
| English | 39.8 | 55.8 | 41.3 | 40.2 | 57.2 | 55.2 | 58.5 | 53.9 |
| Finnish | 26.2 | 27.7 | 27.7 | 28.3 | 40.3 | 32.5 | 34.3 | 40.4 |
| French | 35.7 | 50.9 | 49.5 | 47.0 | 44.2 | 55.8 | 54.6 | 42.1 |
| German | 49.7 | 47.1 | 56.0 | 51.2 | 49.5 | 55.7 | 57.4 | 49.9 |
| Greek | 61.7 | 70.0 | 62.1 | 60.2 | 60.5 | 68.8 | 62.0 | 60.2 |
| Hebrew | 52.9 | 58.7 | 60.9 | 57.5 | 54.8 | 62.6 | 54.2 | 57.2 |
| Hungarian | 68.8 | 41.6 | 71.3 | 63.6 | 69.2 | 65.5 | 72.4 | 64.8 |
| Indonesian | 32.0 | 58.3 | 58.1 | 43.6 | 50.2 | 58.6 | 58.5 | 59.4 |
| Irish | 63.1 | 64.5 | 65.2 | 63.0 | 63.4 | 64.4 | 64.7 | 63.9 |
| Italian | 62.7 | 77.1 | 73.6 | 72.5 | 69.2 | 65.2 | 69.8 | 72.4 |
| Japanese | 56.4 | 70.5 | 56.9 | 73.9 | 56.9 | 69.0 | 57.0 | 73.5 |
| Persian | 46.9 | 45.1 | 51.2 | 39.7 | 48.0 | 45.1 | 51.1 | 41.7 |
| Spanish | 46.8 | 56.1 | 57.3 | 63.1 | 57.7 | 56.2 | 58.5 | 62.3 |
| Swedish | 43.5 | 44.8 | 43.2 | 43.5 | 57.9 | 53.3 | 53.3 | 56.9 |
| Avg. | 49.3 | 55.2 | 55.3 | 52.9 | 56.7 | 57.6 | 58.2 | 56.9 |

Table 5.4: Accuracy comparison on UD15 for selected configurations with the hard constraints on possible root POS tags.

or a noun or a adjective if the main verb is copula. We remove adjective from this set as we found it is relative rare across languages. Interestingly in this case, the stack depth constraints ($C = 2$ and $C = 3$) work the best. In particular, the average score of the length bias ($\beta_{len} = 0.1$) drops. We inspect the reason of this below.

We next see the effect of another constraint, the verb-otherwise-noun constraint, which excludes nouns from the candidate for the root if both a verb and noun exist. This probably decreases the recall though we expect that it increases the performance as the majority of predicates is verbs. As we expected, with this constraint the average performance of baseline uniform model increases sharply from 49.2 to 56.7 (+7.5), which is larger than any increases with structural constraint to the original baseline model. In this case, though the change is small, again our stack depth constraints perform the best (58.2 with $C = 3$); the average score with the length bias does not increase.

5.4.2 Qualitative analysis

When we inspect the scores of the models without root POS constraints in Table 5.3 and the models with the verb-or-noun constraint in Table 5.4, we notice that the behaviors of our models with stack

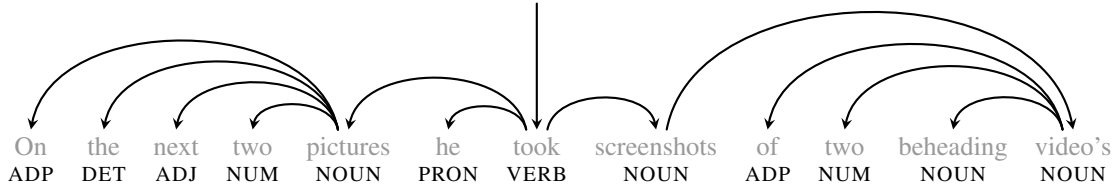
depth constraints and other models are often quite different. Specifically,

- It is only Greek on which the baseline uniform model improves score from the setting with no root POS constraint.
- In other languages, the scores of the uniform model are unchanged or dropped when adding the root POS constraint. For example the score for Czech drops from 61.8 to 50.7.
- The same tendency is observed in the models of $\beta_{len} = 0.1$. Its score for Greek improves from 30.4 to 60.2 while other scores are often unchanged or dropped; an exception is French, on which the score improves from 35.7 to 47.0. For other languages, such as Croatian (-7.2), English (-16.4), Indonesian (-6.4), and Swedish (-12.4), the scores sharply drop.
- On the other hand, we observe no significant performance drops in our models with stack depth constraints (i.e., $C = 2$ and $C = 3$) by adding the root POS constraint.

These differences between the length constraints and stack depth constraints are interesting and may shed some light on the characteristics of two approaches. Here we look into the output parses in English, with which the performance changes are typical, i.e., the model of $\beta_{len} = 0.1$ drops while the scores of other models are almost unchanged.

When we compare output parses of different models, we notice that often the same tree is predicted by several different models. Figure 5.6 shows examples of output parses of different models. The following observations are made with those errors. Note that these are typical, in that the same observation can be often made on other sentences as well.

1. One strong observation from Figure 5.6 is that the output of $\beta_{len} = 0.1$ reduces to that of the uniform model when the root POS constraint is added to the model. As can be seen in other parses, every model in fact predicts that the root token is a noun or a verb, which suggests this explicit root POS constraint is completely redundant in the case of English.
2. Contrary to $\beta_{len} = 0.1$, the stack depth constraints, $C = 2$ and $C = 3$, are not affected by the root POS constraint. This is consistent with the scores in Tables 5.3 and 5.4; the score of $C = 3$ is unchanged and that of $C = 2$ increases by just 1.0 point with the root POS constraint.
3. While the scores of the uniform model and $C = 3$ are similar in Table 5.3 (38.7 and 41.3, respectively), the properties of output parses seem very different. The typical errors made by $C = 3$ are the root tokens, which are in most cases predicted as nouns as in Figure 5.6(d), and arcs between nouns and verbs, which also are typically predicted as $\text{NOUN} \rightarrow \text{VERB}$. Contrary to these *local* mistakes, the uniform model often fails to capture the basic structure of a sentence. For example, while $C = 3$ correctly identifies that “of two beheading video’s” comprises a constituent, which modifies “screenshots”, which in turn becomes an argument of “took”, the parse of the uniform model is more corrupt in that we cannot identify any semantically coherent units from it. See also Figure 5.7 where we compare outputs of these models on another sentence.



(a) Gold parse.

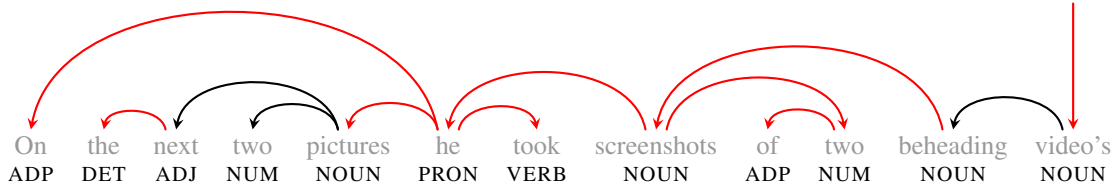
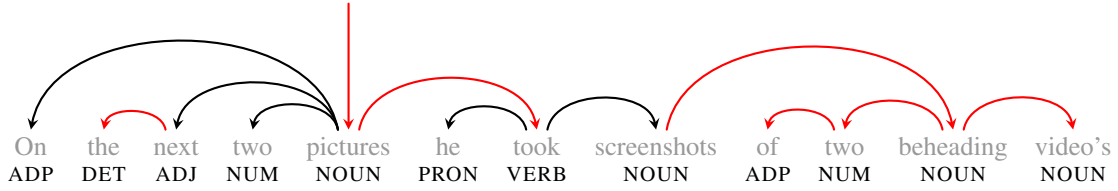
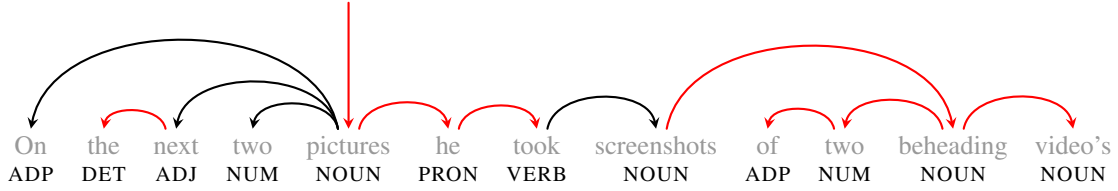
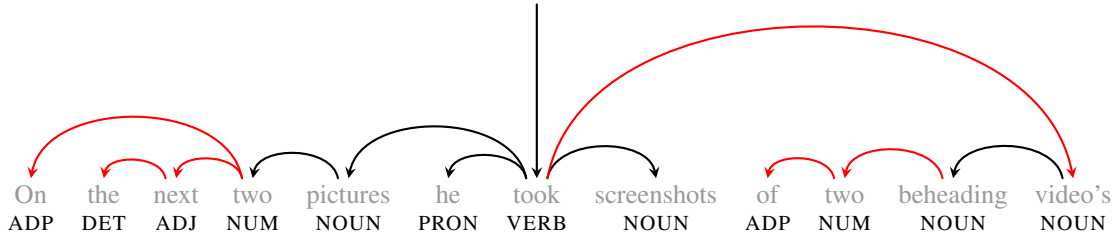
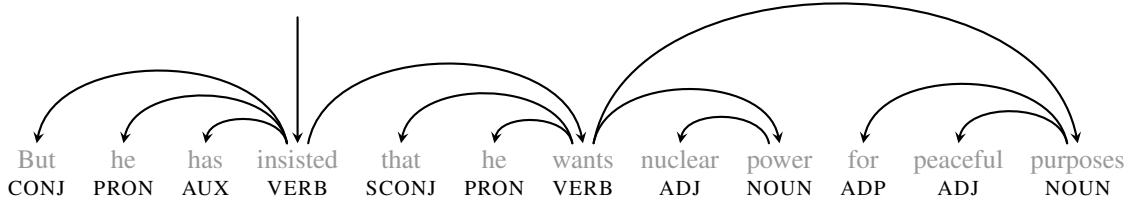
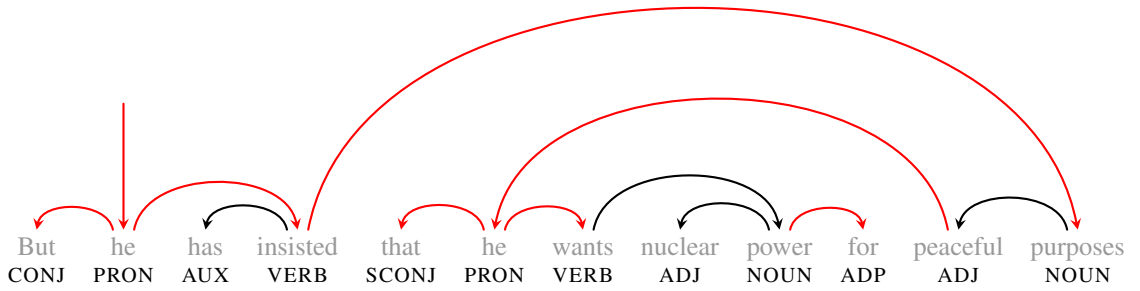
(b) Output by uniform, uniform + verb-or-noun, and $\beta = 0.1 + \text{verb-or-noun}$.(c) Output by $C = 2$, $C = 2 + \text{verb-or-noun}$.(d) Output by $C = 3$, $C = 3 + \text{verb-or-noun}$.(e) Output by $\beta_{len} = 0.1$.

Figure 5.6: Comparison of output parses of several models on a sentence in English UD. The outputs of $C = 2$ and $C = 3$ do not change with the root POS constraint, while the output of $\beta_{len} = 0.1$ changes to the same one of the uniform model with the root POS constraint. Colored arcs indicate the wrong predictions. Note surface forms are not observed by the models (only POS tags are).



(a) Gold parse.



(b) Output by the uniform model.

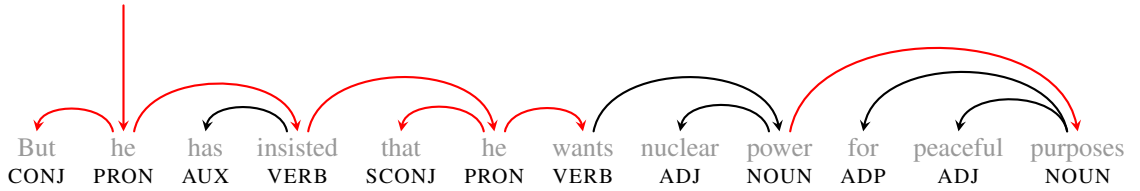
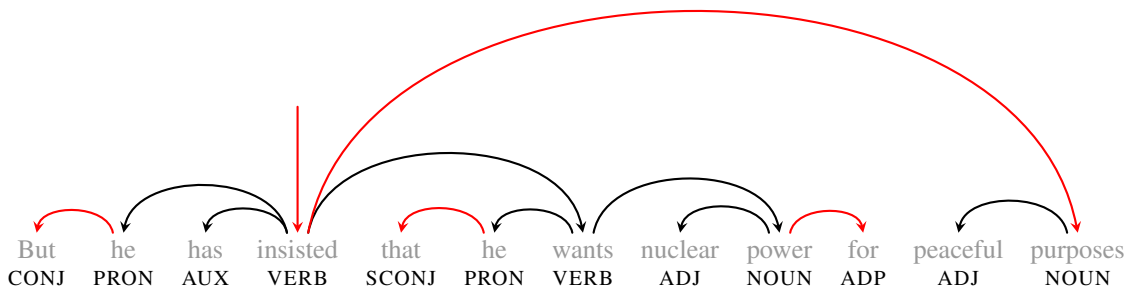
(c) Output by $C = 3$.(d) Output by $\beta_{len} = 0.1$.

Figure 5.7: Another comparison between outputs of the uniform model and $C = 3$ in English UD. We also show $\beta_{len} = 0.1$ for comparison. Although the score difference is small (see Table 5.3), the types of errors are different. In particular the most of parse errors by $C = 3$ are at local attachments (first-order). For example it consistently recognizes a noun is a head of a verb, and a noun is a sentence root. Note an error on “power \rightarrow purposes” is an example of PP attachment errors, which may not be solved under the current problem setting receiving only a POS tag sequence.

Discussion The first observation, i.e., the output of $\beta_{len} = 0.1 + \text{verb-or-noun}$ reduces to that of the uniform model, is found in most other sentences as well. Along with the results in other languages, we suspect the effect of the length bias gets weak when the root POS constraint is given. We do not analyze the cause of this degradation more, but the discussion below on the difference between two constraints, i.e., the stack depth constraint and the length bias, might be relevant to that.

The essential difference between these two approaches is in the assumed structural form to be constrained: The length bias (i.e., β_{len}) is a bias for each dependency arcs on the tree, while the stack depth constraint, which corresponds to the center-embeddedness, is inherently the constraint on constituent structures. Interestingly, we can see the effect of this difference in the output parses in Figures 5.6 and 5.7. Note that we do not use the constraints at decoding and all differences are due to the learned parameters with the constraints during training.

Nevertheless, we can detect some typical errors in two approaches. One difference between trees in Figure 5.6 is in the constructions of a phrase “On ... pictures”. $\beta_{len} = 0.1$ predicts that “On the next two” comprises a constituent, which modifies “pictures” while $C = 2$ and $C = 3$ predict that “the next two pictures” comprises a constituent, which is correct, although the head of a determiner is incorrectly predicted. On the other hand, $\beta_{len} = 0.1$ works well to find more primitive dependency arcs between POS tags, such as arcs from verbs to nouns, which are often incorrectly recognized by stack depth constraints. Similar observations can be made in trees in Figure 5.7. See the constructions on “for peaceful purposes”. In is only $C = 3$ (and $C = 2$ though we omit) that predicts it becomes a constituent. In other positions, again, $\beta_{len} = 0.1$ works better to find local dependency relationships. The head of “purposes” is predicted differently, but this choice is equally difficult in the current problem setting (see the caption of Figure 5.7).

These observations may explain the reason why the root POS constraints work better with the stack depth constraints than the dependency length bias. With the stack depth constraints, the main source of improvements is detections of constituents, but this constraint itself does not help to resolve some dependency relationships, e.g., the dependency direction between verbs and nouns. The root POS constraints are thus orthogonal to this approach. They may help to solve the remaining ambiguities, e.g., the head choice between a noun and a verb. On the other hand, the dependency length bias is the most effective to find basic dependency relationships between POS tags while the resulting tree may contain implausible constituent units. Thus the effect of the length bias seems somewhat overlapped with the root POS constraints, which may be the reason why they do not well collaborate with each other.

Other languages We further inspect the results of some languages with exceptional behaviors separately below.

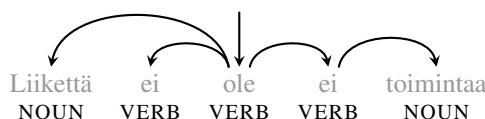
Japanese In Figure 5.5, we can see that the performance of Japanese is the best with a strong stack depth constraint, such as depth 1 and $C = 2$, and the performance drops when relaxing the constraint. This may be counterintuitive from our oracle results in Chapter 4 (e.g., Figure 4.13) that Japanese is the language in which the ratio of center-embedding is relatively higher.

Inspecting the output parses, we found that these results are essentially due to the word order of Japanese, which is mainly head final. With a strong constraint (e.g., the stack depth

one), the model tries to build a parse that is purely left- or right-branching. An easy way to create such parse is placing a root word at the beginning or the end of the sentence, and then connecting adjacent tokens from left to right, or right to left. This is what happened when a severe constraint, e.g., the maximum stack depth of 1 is imposed. Since the position of root token is in most cases correctly identified, the score gets relatively higher. On the other hand, when relaxing the constraint, the model also try to explore parses in which the root token is not the beginning/end of the sentence, but internal positions, and the model fail to find the head final pattern of Japanese.

This Japanese result suggests that sometimes our stack depth constraint helps learning even when the imposed stack depth bound does not fit well to the syntax of the target language, though the learning behavior differs from our expectation. In this case, the model does not capture the syntax correctly in the sense that Japanese sentences cannot be parsed with a severe stack depth bound, but the model succeeded to find syntactic patterns that are a very rough approximation of the true syntax, resulting in a higher score.

Finnish Finnish is an inflectional language with rich morphologies and with little function words. This is essentially the reason for consistent lower accuracies of Finnish even when the constraint on root POS tags is given. Recall that all our models are imposed the function word constraint (Section 5.3.4). Though our primary motivation to introduce this constraint is to alleviate problems in evaluation, it also greatly reduces the search space if the ratio of function words is high. Also at test time, a higher ratio of function words indicates a higher chance of correct attachments since the head candidates for a function word is limited to other content words.⁵ Below is an example of a dependency tree in Finnish treebank:



This sentence comprises of NOUN and VERB only, and there are a lot of similar sentences. This example also explains the reason why the performance of Finnish is still low with the root POS constraints. Table 5.5 lists the statistics about the ratio of function words in the training corpora. We can see that it is only Finnish that the ratio of function words is less than 10%. Also, the ratio in Japanese is very high. This probably explains the reason for relatively high overall scores of Japanese. Thus, the variation of the scores across languages in the current experiment is largely explained by the ratio of function words in each language.

Greek In Figure 5.3, the scores on Greek with the stack depth constraints are consistently worse than the uniform baseline. Though overall scores are low, the situation largely changes with the root POS constraints, and with them the scores get stable.

⁵Recall that although we remove constraints at test time the model rarely find a parse with function words at internal positions since the model is trained to avoid such parses.

| | Ratio (%) | | Ratio (%) |
|-----------|-------------|------------|--------------|
| basque | 26.57 | hebrew | 32.29 |
| bulgarian | 25.88 | hungarian | 23.76 |
| croatian | 24.55 | indonesian | 19.68 |
| czech | 20.09 | irish | 36.09 |
| danish | 30.66 | italian | 37.73 |
| english | 27.98 | japanese | 45.14 |
| finnish | 9.63 | persian | 23.25 |
| french | 37.84 | spanish | 36.99 |
| german | 32.09 | swedish | 29.64 |
| greek | 16.94 | | |

Table 5.5: Ratio of function words in the training corpora of UD (sentences of length 15 or less).

A possible explanation for these exceptional behaviors might be the relatively small ratio of function words (Table 5.5) in the data along with the small size of the training data (Table 5.1), both of which could be partially alleviated with the root POS constraints.

More linguistically intuitive explanation might be that Greek is a relatively free word order language and our structural constraints do not work well for guiding the model for finding such grammars. However, to make such conclusion, we have to set up experiments more carefully, e.g., by eliminating the bias caused by the smaller size of the data. We thus leave it our future work to discuss the limitation of the current approach with a typological difference in each language.

5.4.3 Google Universal Dependency Treebanks

So far our comparison is limited in the models of our baseline DMV model with some constraints. Next we see the relative performance of this approach compared to the current state-of-the-art unsupervised systems on another dataset, Google treebanks.

Table 5.6 shows the result. The scores of the other systems are borrowed from Grave and Elhadad (2015). In this experiment, we only focus on the settings where the root word is restricted with the verb-otherwise-noun constraint. Among our structural constraints, again our stack depth constraints perform the best. In particular the scores with $C = 2$ are stable across languages.

All our method outperforms the strong baseline model of Naseem et al. (2010), which encodes manually crafted rules (12 rules) such as $\text{VERB} \rightarrow \text{NOUN}$ and $\text{ADP} \rightarrow \text{NOUN}$ via the posterior regularization method (Ganchev et al., 2010). Compared to this, our baseline method uses fewer syntactic rules via parameter-based constraints, in total 5 (3 for function words and the verb-otherwise-noun constraint) and is much simpler than their posterior regularization method.

Grave and Elhadad (2015) is a more sophisticated model, which utilizes the same syntactic rules as the Naseem et al.’s method. Our models do not outperform this model, though it is only Korean that ours do not perform competitively to their model.

| | Unif. | $C = 2$ | $C = 3$ | $\beta_{len} = 0.1$ | Naseem10 | Grave15 |
|---------------|-------|---------|---------|---------------------|----------|---------|
| German | 64.5 | 64.3 | 64.6 | 62.5 | 53.4 | 60.2 |
| English | 57.9 | 59.5 | 57.9 | 56.9 | 66.2 | 62.3 |
| Spanish | 68.2 | 71.1 | 70.5 | 69.6 | 71.5 | 68.8 |
| French | 69.2 | 69.6 | 70.1 | 66.4 | 54.1 | 72.3 |
| Indonesian | 66.8 | 67.4 | 66.0 | 66.7 | 50.3 | 69.7 |
| Italian | 43.9 | 67.3 | 65.9 | 44.0 | 46.5 | 64.3 |
| Japanese | 47.5 | 54.5 | 47.4 | 47.6 | 58.2 | 57.5 |
| Korean | 28.6 | 30.7 | 28.3 | 43.2 | 48.8 | 59.0 |
| Br-Portuguese | 63.0 | 67.1 | 62.7 | 62.6 | 46.4 | 68.3 |
| Swedish | 67.4 | 67.9 | 67.3 | 66.4 | 64.3 | 66.2 |
| Avg | 57.7 | 62.0 | 60.1 | 58.6 | 56.0 | 64.8 |

Table 5.6: Attachment scores on Google universal treebanks (up to length 10). All proposed models are trained with the verb-otherwise-noun constraint. Naseem10 = the model with manually crafted syntactic rules between POS tags (Naseem et al., 2010); Grave15 = also relies on the syntactic rules but is trained discriminatively (Grave and Elhadad, 2015).

5.5 Discussion

We found that our stack depth constraints improve the performance of unsupervised grammar induction across languages and datasets in particular when some seed knowledge about grammar is given to the model. However, we also find that in many languages the improvements from the no structural constraint baseline becomes small when such knowledge is given. Also, the performance reaches to the current state-of-the-art method, which utilizes much more complex machine learning techniques as well as manually specific syntactic rules. We thus speculate that our models already reach some limitation under the current problem setting, that is, learning of dependency grammar from the POS input only.

Recall that the annotation of UD is content-head based and every function word is a dependent of the most closely related content word. This means under the current first-order model on POS tag inputs, many important information that currently a supervised parser would exploit is abandoned. For example, the model would collapses both some noun phrase and prepositional phrase into its head (probably NOUN) while this information is crucial; e.g., an adjective cannot be attached to a prepositional phrase, etc. One way to exploit such clue for disambiguation is to utilize the boundary information (Spitkovsky et al., 2012; Spitkovsky et al., 2013).

Typically, unsupervised learning of structures gets more challenging when employing more structurally complex models. However, one of our strong observations from the current experiment is that our stack depth constraint rarely hinders, i.e., does not decrease the performance. Here we have focused on very simple generative model of DMV though it may be more interesting to see what happens when imposing this constraint on more structurally complex models on which learning is much harder. There remains many rooms for further improvements and such type of study would be important toward one of the goals of unsupervised parsing of identifying which

structure can be learned without explicit supervisions (Bisk and Hockenmaier, 2015).

In this work, our main focus was the imposed structural constraints (or linguistic prior) themselves, and we did not care much about the method to encode these prior knowledge. That is, our method to inject constraints was a crude way, i.e., via hard constraints in the E step (Eq. 5.2), and there exist more sophisticated methods with newer machine learning techniques. Posterior regularization (PR) that we compared the performance with (Naseem et al., 2010) is one of such techniques. The crucial difference between our imposing hard constraints and PR is that PR imposes constraints in expectation, i.e., every constraint becomes a *soft* constraint.

If we reformulate our model with PR, then that may probably impose a soft constraint, e.g., *the expected number of occurrence of center-embedding up to some degree in a parse is less than X*. X becomes 1 if we try to resemble the behavior of our hard constraints, but could be any values, and such flexibility is one advantage, which our current method cannot appreciate. Thus, from a machine learning perspective, one interesting direction is to compare the performances of two approaches with the (conceptually) same constraints.

5.6 Conclusion

In this study, we have shown that our imposed stack depth constraint improves the performance of unsupervised grammar induction in many settings. Specifically, it often does not harm the performance when it already performs well while it reinforces the relatively poorly performed models (Table 5.6). One limitation of the current approach is that the information that the parser can utilize is very superficial (i.e., the first order model on content-head based POS tags). However, our positive results in the current experiment are an important first step for the current line of research and encourage further study on more structurally complex model beyond the simple DMV model.

Chapter 6

Conclusions

Identifying universal syntactic constraints of language is an attractive goal both from the theoretical and empirical viewpoints. To shed light on this fundamental problem, in this thesis, we pursued the *universality* of the language phenomena of center-embedding avoidance, and its practical utility in natural language processing tasks, in particular unsupervised grammar induction. Along with these investigations, we develop several computational tools capturing the syntactic regularities stemming from center-embedding.

The tools we presented in this thesis are left-corner parsing methods for dependency grammars. We formalized two related parsing algorithms. The transition-based algorithm presented in Chapter 4 is an incremental algorithm, which operates on the stack, and its stack depth only grows when processing center-embedded constructions. We then considered tabulation of this incremental algorithm in Chapter 5, and obtained an efficient polynomial time algorithm with the left-corner strategy. In doing so, we applied *head-splitting* techniques (Eisner and Satta, 1999; Eisner, 2000), with which we removed the spurious ambiguity and reduced the time complexity from $O(n^6)$ to $O(n^4)$, both of which were essential for our application of inside-outside calculation with filtering.

Dependency grammars were the most suitable choice for our cross-linguistic analysis on language universality, and we obtained the following fruitful empirical findings using the developed tools for them.

- Using multilingual dependency treebanks, we quantitatively demonstrate the universality of center-embedding avoidance. We found that most syntactic constructions across languages can be covered within highly restricted bounds on the degree of center-embedding, such as one, or zero, when relaxing the condition of the size of embedded constituent.
- From the perspective of parsing *algorithms*, the above findings mean that a left-corner parser can be utilized as a tool for exploiting universal constraints during parsing. We verified this ability of the parser empirically by comparing the growth of stack depth when analyzing sentences on treebanks with those of existing algorithms, and showed that only the behavior of the left-corner parser is consistent across languages.
- Based on these observations, we examined whether the found syntactic constraints help in

finding the syntactic patterns (grammars) in the given sentences through experiments on unsupervised grammar induction, and found that our method often boosts the performance from the baseline, and competes with the current state-of-the-art method in a number of languages.

We believe the presented study will be the starting point of many future inquiries. As we have mentioned several times, our choice of dependency grammars for the representation was motivated by its cross-linguistic suitability as well as its computational tractability. Now we have evidences on the language universality of center-embedding avoidance. We consider thus one exciting direction would be to explore unsupervised learning of constituent structures exploiting our found constraint, which has not been solved yet with traditional PCFG-based methods. Note that unlike the dependency length bias, which is only applicable for dependency-based models, our constraint is conceptually free from grammar formalisms.

As we have mentioned in Chapter 1, recently there has been a growing interest on the tasks of grounding, or semantic parsing. Also, another direction of grammar induction in particular with more sophisticated grammar formalisms, such as CCG, has been initiated with some success. There remains many open questions in these settings, e.g., on the necessary amount of seed knowledge to make learning tractable (Bisk and Hockenmaier, 2015; Garrette et al., 2015). Arguably, the system with less initial seed knowledge or less assumption on the specific task is preferable (by keeping accuracies). We hope our introduced constraint helps in reducing the required assumption, or improving the performance in those more general grammar induction tasks. Finally, we suspect that the study of child language acquisition would also have to be discussed within the setting of grounding, i.e., with some kind of distant supervision or perception. Although we have not explored the cognitive plausibility of the presented learning and parsing methods, our empirical finding that when learning from relatively short sentences a severe stack depth constraint (relaxed depth one) often improves the performance may become an appealing starting point for exploring computational models of child language acquisition with human-like memory constraints.

Appendix A

Analysis of Left-corner PDA

This appendix contains the proof of Theorem 2.1, which establishes the connection between the stack depth of the left-corner PDA and the degree of center-embedding. For proving this, we first need to extend the notion of center-embedding for a *token* as follows:

Definition A.1. *Given a sentence and token e (not the initial token) in the sentence, we write the derivation from S to e as follows with the minimal number of \Rightarrow :*

$$\begin{aligned}
 S &\Rightarrow_{lm}^* v \underline{A} \alpha \Rightarrow_{lm}^+ v w_1 \underline{B_1} \alpha \Rightarrow_{lm}^+ v w_1 \underline{C_1} \beta_1 \alpha \\
 &\Rightarrow_{lm}^+ v w_1 w_2 \underline{B_2} \beta_1 \alpha \Rightarrow_{lm}^+ v w_1 w_2 \underline{C_2} \beta_2 \beta_1 \alpha \\
 &\Rightarrow_{lm}^+ \dots \\
 &\Rightarrow_{lm}^+ v w_1 \dots w_{m_e} \underline{B_{m_e}} \beta_{m_e-1} \dots \beta_1 \alpha \Rightarrow_{lm}^* v w_1 \dots w_{m_e} \underline{C_{m_e}} \beta_{m_e} \beta_{m_e-1} \dots \beta_1 \alpha \\
 &\Rightarrow_{lm}^* v w_1 \dots w_{m_e} \underline{x' E} \beta_{m_e} \beta_{m_e-1} \dots \beta_1 \alpha \Rightarrow_{lm} v w_1 \dots w_{m_e} \underline{x' e} \beta_{m_e} \beta_{m_e-1} \dots \beta_1 \alpha,
 \end{aligned} \tag{A.1}$$

where the underlined symbol is the expanded symbol by the following \Rightarrow . Then, the degree of center-embedding for token e is:

- $m_e - 1$ if $C_{m_e} = E$ (i.e., $x' = \varepsilon$) or $B_{m_e} = C_{m_e} = E$ (i.e., $\beta_{m_e} = x' = \varepsilon$); and
- m_e otherwise.

The degree of the token at the beginning of the sentence is defined as 0.

The main difference of Eq. A.1 in this definition from Eq. 2.1 is that instead of expanding C_{m_e} to string x , we take into account the right edges from C_{m_e} to another nonterminal (preterminal) E , which should exist if the requisite in Eq. 2.1 that $|x| \geq 2$ is satisfied. Eq. A.1 explains a zig-zag path from the start symbol S to a token (terminal e), which can be classified into three cases in Figure A.1. Definition A.1 determines the degree of center-embedding of that token depending on the structure of this path, which will be explained further below.

- Given terminal e , the derivation of the form in Eq. A.1 is deterministic, and each B_i or C_i is determined as a turning point on a zig-zag path; see e.g., a path from c to S in Figure 2.10(c). A is the starting point, which might be identical to S . This is indicated with dotted edges in Figure A.1; Figure 2.10(c) is such a case.

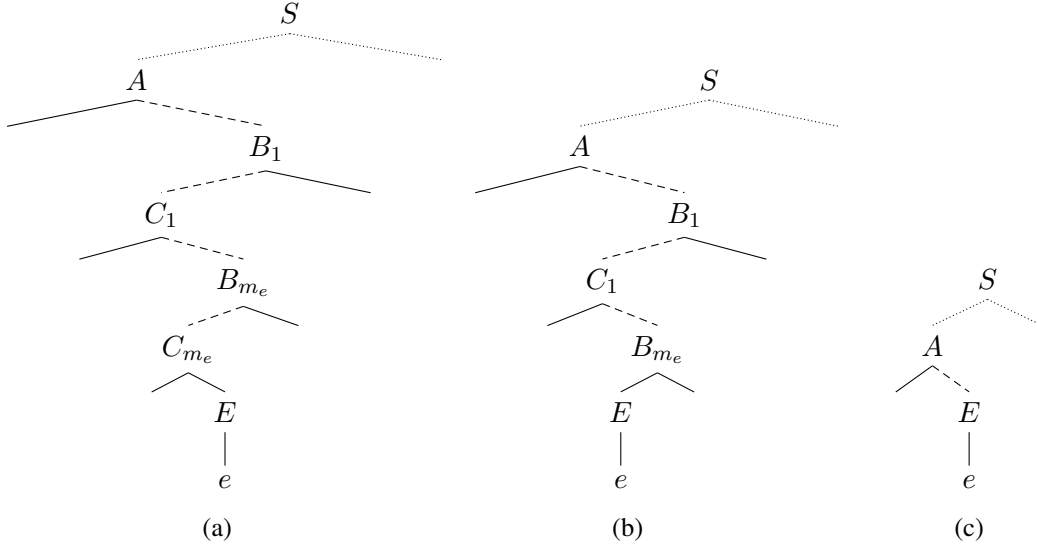


Figure A.1: Three types of realizations of Eq. A.1. Dashed edges may consist of more than one edge (see Figure A.2 for example) while dotted edges may not exist (or consist of more than one edge). (a) E is a right child of C_{m_e} and thus the degree of center-embedding is m_e . (b) E is a left child of B_{m_e} (i.e., $C_{m_e} = E$) and the degree is $m_e - 1$; when $m = 1$, $C_1 = E$ and thus no center-embedding occurs. (c) $B_{m_e} = C_{m_e} = E$; note this happens only when $m_e = 1$ (see body).

- We allow an empty transition from B_{m_e} to C_{m_e} and C_{m_e} to E at the last transitions in Eq. A.1, which are important to define the degree in the case where the preterminal E for token e is not the right child of the parent ($C_{m_e} = E$), or no center-embedding is involved (i.e., $B_{m_e} = C_{m_e} = E$ and $m_e = 1$). Figure A.1(b) is the complete case without empty transitions, while Figures A.1(b) and A.1(c) involve empty transitions. Figure A.1(b) with $m_e = 1$, where the degree is $m_e - 1 = 0$, is an example of the parse in Figure 2.10(d), where no center-embedding is involved (b corresponds to e in Figure A.1(b)). Figure A.1(c) is the case where the empty transition from B_{m_e} to C_{m_e} occurs. Note that this pattern only occurs for $m_e = 1$, which includes the derivation to the last token of the sentence, where the path is always right edges from S (or A) to B_1 (or E) and the degree is 0. This is because the derivation with an empty transition from B_{m_e} to C_{m_e} indicates $B_{m_e} = E$, though when $m > 1$, it safely reduces to the case of $m_e - 1$ in Figure A.1(a).
- Given a CFG parse, the maximum value of the degree in Definition A.1 among tokens in the sentence is identical to the degree of center-embedding defined for that parse (Definition 2.2).

We next prove the following lemma, which is closely connected to Theorem 2.1.

Lemma A.1. *Given token e (not the initial token) in the sentence, let m'_e be the degree of center-embedding of it, and δ_e be the stack depth before it is shifted for recognizing that parse on the left-corner PDA. Then, $\delta_e = m'_e + 1$.*

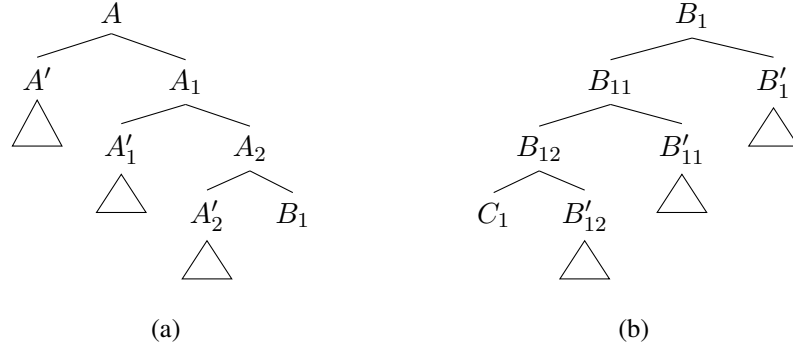


Figure A.2: (a) Example of realization of a path between A and B_1 in Figure A.1. (b) The one between B_1 and C_1 .

Proof. The path from S to every token in the sentence except the beginning of the sentence can be classified into three cases in Figure A.1. We show that in every case between the stack depth δ_e before e is shifted and the degree of center-embedding m'_e , $\delta_e = m'_e + 1$ holds.

Note first that in all cases, the existence of edges from S to A (i.e., whether $S = A$ or not) does not affect the stack depth δ_e . This is due to the basic order of building a parse in the left-corner PDA, which always completes a left subtree first, and then expands it with PREDICTION. Thus, in the following, we ignore the existence of S , and focus on the stack depth at e during building a subtree rooted at A .

- (a) The path from C_{m_e} to E exists (Figure A.1(a)): In this case, the degree of center-embedding $m'_e = m_e$. Before shifting e , the following stack configuration occurs:

$$A/B_1 \underbrace{C_1/B_2 \ C_2/B_3 \ \cdots \ C_{m_e-1}/B_{m_e} \ C_{m_e}/E}_{m_e}, \quad (\text{A.2})$$

This can be shown as follows.

The PDA first makes symbol A/B_1 . Note that the path from A to B_1 may contain many nonterminals as shown in Figure A.2(a). During processing these nodes, the PDA first builds a subtree rooted at A' , then performs PREDICTION, which results in symbol A/A_1 . After that, A'_1 is built with A/A_1 being remained on the stack, and then connect them with COMPOSITION, which results in A/A_2 . Finally A/B_1 remains on the stack after repeating this process.

Then, C_1/B_2 is made on the stack in the similar manner, but both symbols remain on the stack, since A/B_1 cannot be combined with another subtree unless it is complete (without a predicted node). There may exist many nonterminals between B_1 and C_1 as in Figure A.2(b), but they does not affect the configuration of the stack; for example, B_{12} is first introduced after a subtree rooted at C_1 is complete. This indicates that the stack accumulates symbols C_i/B_{i+1} as the number of right edges between them increases. Finally, after building a subtree rooted at the left child of C_{m_e} , it is converted to C_{m_e}/E by PREDICTION, resulting in

the stack configuration of Eq. A.2. This occurs just before e is shifted on the stack by SCAN.

- (b) $C_{m_e} = E$ (Figure A.1(b)): In this case $m'_e = m_e - 1$. Before shifting e , the stack configuration is:

$$A/B_1 \underbrace{C_1/B_2 \ C_2/B_3 \ \cdots \ C_{m_e-1}/B_{m_e}}_{m_e-1}. \quad (\text{A.3})$$

e is shifted on this stack by SHIFT, and then COMPOSITION is performed between C_{m_e-1}/B_{m_e} and E . Thus $\delta_e = m_e = m'_e + 1$.

- (c) $B_1 = C_1 = E$ (Figure A.1(c)): The stack configuration before shifting e is apparently A/E . $m' = 0$, so $\delta_e = m' + 1$ holds.

■

Now the proof of Theorem 2.1 is immediate from Lemma A.1.

Proof of Theorem 2.1. The relationship between Definitions 2.2 and A.1 is that the maximum value of the degree given by Definition A.1 for each token is the same as the degree of a parse. Given e, δ_e, m'_e in Lemma A.1, $\delta_e = m'_e + 1$. Let $e^* = \arg \max_e \delta_e$. “The maximum value of the stack depth after a reduce transition” in Theorem 2.1 can be translated to the maximum value *before* a reduce transition, which is δ_{e^*} . Thus, $\delta_{e^*} = m'_{e^*} + 1$. Arranging, $m'_{e^*} = \delta_{e^*} - 1$. ■

Appendix B

Part-of-speech tagset in Universal Dependencies

Universal Dependencies (UD) uses the following 17 part-of-speech (POS) tags.

- ADJ: adjective
- ADP: adposition
- ADV: adverb
- AUX: auxiliary verb
- CONJ: coordinating conjunction
- DET: determiner
- INTJ: interjection
- NOUN: noun
- NUM: numeral
- PRON: pronoun
- VERB: verb
- PART: particle
- PRON: pronoun
- SCONJ: subordinating conjunction
- PUNCT: punctuation
- SYM: symbol
- X: other

Bibliography

- Steven Abney and Mark Johnson. 1991. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, 20(3):233–250.
- Stephen P. Abney. 1987. *The English Noun Phrase In Its Sentential Aspect*. Ph.D. thesis, M.I.T.
- Itzair Aduriz, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Aitzpea Garmendia, and Maite Oronoz. 2003. Construction of a Basque dependency treebank. In *Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories*.
- Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. 2002. “Floresta sintá(c)tica”: a treebank for Portuguese. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, pages 1698–1703, Las Palmas, Spain.
- Željko Agić and Nikola Ljubešić. 2014. The SETimes.HR linguistically annotated corpus of Croatian. In *Proceedings of LREC 2014*, pages 1724–1727, Reykjavík, Iceland.
- Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation, and Compiling*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.
- Hiyan Alshawi. 1996. Head automata and bilingual tiling: Translation with minimal representations (invited talk). In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Santa Cruz, California, USA, June. Association for Computational Linguistics.
- Waleed Ammar, Chris Dyer, and Noah A Smith. 2014. Conditional random field autoencoders for unsupervised structured prediction. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3311–3319. Curran Associates, Inc.
- Nart B. Atalay, Kemal Oflazer, and Bilge Say. 2003.
- Miguel Ballesteros and Joakim Nivre. 2013. Going to the roots of dependency parsing. *Computational Linguistics*, 39(1):5–13.
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in*

- Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. 2010. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, California, June. Association for Computational Linguistics.
- Niels Beuck and Wolfgang Menzel. 2013. Structural prediction in incremental dependency parsing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 7816 of *Lecture Notes in Computer Science*, pages 245–257. Springer Berlin Heidelberg.
- Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.
- Yonatan Bisk and Julia Hockenmaier. 2012. Simple robust grammar induction with combinatory categorial grammars. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, July 22-26, 2012, Toronto, Ontario, Canada*.
- Yonatan Bisk and Julia Hockenmaier. 2013. An hdp model for inducing combinatory categorial grammars. *Transactions of the Association for Computational Linguistics*, 1:75–88.
- Yonatan Bisk and Julia Hockenmaier. 2015. Probing the linguistic strengths and limitations of unsupervised grammar induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1395–1404, Beijing, China, July. Association for Computational Linguistics.
- Yonatan Bisk, Christos Christodoulopoulos, and Julia Hockenmaier. 2015. Labeled grammar induction with minimal supervision. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 870–876, Beijing, China, July.
- Phil Blunsom and Trevor Cohn. 2010. Unsupervised induction of tree substitution grammars for dependency parsing. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1204–1213, Cambridge, MA, October. Association for Computational Linguistics.
- Bernd Bohnet, Joakim Nivre, Igor M. Boguslavsky, Richárd Farkas, Filip Ginter, and Jan Hajič. 2013. Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Association for Computational Linguistics*, 1(Oct):429–440.
- Sabine Brants, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

- Sabine Buchholz and Erwin Marsi. 2006. Conll-x shared task on multilingual dependency parsing. In *Proceedings of the Tenth Conference on Computational Natural Language Learning (CoNLL-X)*, pages 149–164, New York City, June. Association for Computational Linguistics.
- Glenn Carroll, Glenn Carroll, Eugene Charniak, and Eugene Charniak. 1992. Two experiments on learning probabilistic dependency grammars from corpora. In *Working Notes of the Workshop Statistically-Based NLP Techniques*, pages 1–13. AAAI.
- Eugene Charniak. 1993. *Statistical language learning*. MIT Press.
- Keh-Jiann Chen, Chi-Ching Luo, Ming-Chung Chang, Feng-Yi Chen, Chao-Jan Chen, Chu-Ren Huang, and Zhao-Ming Gao. 2003. Sinica treebank. In Anne Abeillé, editor, *Treebanks*, volume 20 of *Text, Speech and Language Technology*, pages 231–248. Springer Netherlands.
- Evan Chen, Edward Gibson, and Florian Wolf. 2005. Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, 52(1):144 – 169.
- Montserrat Civit and MaAntònia Martí. 2004. Building cast3lb: A spanish treebank. *Research on Language and Computation*, 2(4):549–574.
- Alexander Clark. 2001. Unsupervised Induction of Stochastic Context-Free Grammars using Distributional Clustering. In *Proceedings of the ACL 2001 Workshop on Computational Natural Language Learning (CoNLL)*.
- Shay Cohen and Noah A. Smith. 2009. Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 74–82, Boulder, Colorado, June. Association for Computational Linguistics.
- Shay Cohen. 2011. *Computational Learning of Probabilistic Grammars in the Unsupervised Setting*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA.
- Trevor Cohn, Phil Blunsom, and Sharon Goldwater. 2010. Inducing tree-substitution grammars. *Journal of Machine Learning Research*, 11:3053–3096, December.
- Michael Collins. 1997. Three Generative, Lexicalised Models for Statistical Parsing. In *35th Annual Meeting of the Association for Computational Linguistics*.
- M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.
- Dóra Csendes, János Csirik, Tibor Gyimóthy, and András Kocsor. 2005. The Szeged treebank. In *TSD*, pages 123–131.
- C. de Marcken. 1999. On the unsupervised induction of phrase-structure grammars. In Susan Armstrong, Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann, and David Yarowsky, editors, *Natural Language Processing Using Very Large Corpora*, volume 11 of *Text, Speech and Language Technology*, pages 191–208. Springer Netherlands.

- Marie-Catherine de Marneffe and Christopher D. Manning. 2008. Stanford typed dependencies manual, September.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38.
- Gabriel Doyle and Roger Levy. 2013. Combining multiple information types in bayesian word segmentation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 117–126, Atlanta, Georgia, June. Association for Computational Linguistics.
- Matthew S. Dryer. 1992. The greenbergian word order correlations. *Language*, 68(1):81–138.
- Sašo Džeroski, Tomaž Erjavec, Nina Ledinek, Petr Pajas, Zdeněk Žabokrtský, and Andreja Žele. 2006. Towards a Slovene dependency treebank. In *Proceedings of the Fifth International Language Resources and Evaluation Conference, LREC 2006*, pages 1388–1391, Genova, Italy. European Language Resources Association (ELRA).
- Jason Eisner and Giorgio Satta. 1999. Efficient parsing for bilexical context-free grammars and head automaton grammars. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL '99*, pages 457–464, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Jason Eisner and Noah A. Smith. 2010. Favor short dependencies: Parsing with soft and hard constraints on dependency length. In Harry Bunt, Paola Merlo, and Joakim Nivre, editors, *Trends in Parsing Technology*, volume 43 of *Text, Speech and Language Technology*, pages 121–150. Springer Netherlands.
- Jason Eisner. 2000. Bilexical Grammars and Their Cubic-Time Parsing Algorithms. In Harry Bunt and Anton Nijholt, editors, *Advances in Probabilistic and Other Parsing Technologies*, pages 29–62. Kluwer Academic Publishers, October.
- Nicholas Evans and Stephen C Levinson. 2009. The myth of language universals: language diversity and its importance for cognitive science. *The Behavioral and brain sciences*, 32(5):429–48; discussion 448–494, October.
- Maryia Fedzechkina, T. Florian Jaeger, and Elissa L. Newport. 2012. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.
- Jenny Rose Finkel, Alex Kleeman, and Christopher D. Manning. 2008. Efficient, feature-based, conditional random field parsing. In *Proceedings of ACL-08: HLT*, pages 959–967, Columbus, Ohio, June. Association for Computational Linguistics.
- Richard Futrell, Kyle Mahowald, and Edward Gibson. 2015. Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences*, 112(33):10336–10341.

- Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. 2010. Posterior regularization for structured latent variable models. *Journal of Machine Learning Research*, 11:2001–2049.
- Dan Garrette, Chris Dyer, Jason Baldridge, and Noah Smith. 2015. Weakly-supervised grammar-informed bayesian ccg parser learning.
- Douwe Gelling, Trevor Cohn, Phil Blunsom, and Joao Graca. 2012. The pascal challenge on grammar induction. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 64–80, Montréal, Canada, June. Association for Computational Linguistics.
- Edward Gibson. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76.
- E. Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In *Image, language, brain: Papers from the first mind articulation project symposium*, pages 95–126.
- Daniel Gildea and David Temperley. 2007. Optimizing grammars for minimum dependency length. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 184–191, Prague, Czech Republic, June. Association for Computational Linguistics.
- Daniel Gildea and David Temperley. 2010. Do grammars minimize dependency length? *Cognitive Science*, 34(2):286–310.
- Kevin Gimpel and Noah A. Smith. 2012. Concavity and initialization for unsupervised dependency parsing. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 577–581, Montréal, Canada, June. Association for Computational Linguistics.
- Yoav Goldberg and Joakim Nivre. 2013. Training deterministic parsers with non-deterministic oracles. *Transactions of the Association for Computational Linguistics*, 1(Oct):403–414.
- Yoav Goldberg. 2011. *Automatic Syntactic Processing of Modern Hebrew (PhD thesis)*.
- Sharon Goldwater, Thomas L. Griffiths, and Mark Johnson. 2009. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21 – 54.
- Carlos Gómez-Rodríguez and Joakim Nivre. 2013. Divisible Transition Systems and Multiplanar Dependency Parsing. *Comput. Linguist.*, 39(4):799–845, December.
- Carlos Gómez-Rodríguez, John A. Carroll, and David J. Weir. 2011. Dependency parsing schemata and mildly non-projective dependency parsing. *Computational Linguistics*, 37(3):541–586.
- Matthew R. Gormley and Jason Eisner. 2013. Nonconvex global optimization for latent-variable models. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 444–454, Sofia, Bulgaria, August. Association for Computational Linguistics.

- Edouard Grave and Noémie Elhadad. 2015. A convex and feature-rich discriminative approach to dependency grammar induction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1375–1384, Beijing, China, July. Association for Computational Linguistics.
- Joseph H. Greenberg. 1963. Some universals of grammar with particular reference to the order of meaningful elements. In Joseph H. Greenberg, editor, *Universals of Human Language*, pages 73–113. MIT Press, Cambridge, Mass.
- Kristina Gulordava, Paola Merlo, and Benoit Crabbé. 2015. Dependency length minimisation effects in short spans: a large-scale analysis of adjective placement in complex noun phrases. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 477–482, Beijing, China, July. Association for Computational Linguistics.
- Aria Haghighi and Dan Klein. 2006. Prototype-driven grammar induction. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 881–888, Sydney, Australia, July. Association for Computational Linguistics.
- Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, Zdeněk Žabokrtský, and Magda Ševčíková-Razímová. 2006. Prague dependency treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia.
- John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *NAACL '01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.
- Katri Haverinen, Timo Viljanen, Veronika Laippala, Samuel Kohonen, Filip Ginter, and Tapio Salakoski. 2010. Treebanking Finnish. In *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories (TLT9)*, pages 79–90.
- Katsuhiko Hayashi, Shuhei Kondo, and Yuji Matsumoto. 2013. Efficient stacked dependency parsing by forest reranking. *Transactions of the Association for Computational Linguistics*, 1:139–150.
- William P. Headden III, Mark Johnson, and David McClosky. 2009. Improving unsupervised dependency parsing with richer contexts and smoothing. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 101–109, Boulder, Colorado, June. Association for Computational Linguistics.

- James Henderson. 2004. Discriminative training of a neural network statistical parser. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 95–102, Barcelona, Spain, July.
- Liang Huang and Kenji Sagae. 2010. Dynamic programming for linear-time incremental parsing. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1077–1086, Uppsala, Sweden, July. Association for Computational Linguistics.
- Liang Huang, Suphan Fayong, and Yang Guo. 2012. Structured perceptron with inexact search. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 142–151, Montréal, Canada, June. Association for Computational Linguistics.
- Richard Hudson. 2004. Are determiners heads? *Functions of Language*, 11(1):7–42.
- T. Florian Jaeger and Harry Tily. 2011. On language utility: Processing complexity and communicative efficiency. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):323–335.
- Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2007. Bayesian inference for PCFGs via Markov chain Monte Carlo. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 139–146, Rochester, New York, April. Association for Computational Linguistics.
- Mark Johnson, Katherine Demuth, and Michael Frank. 2012. Exploiting social information in grounded language learning via grammatical reduction. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 883–891, Jeju Island, Korea, July. Association for Computational Linguistics.
- P. N. Johnson-Laird. 1983. *Mental models: towards a cognitive science of language, inference, and consciousness*. Harvard University Press, Cambridge, MA, USA.
- Mark Johnson. 1998a. Finite-state approximation of constraint-based grammars using left-corner grammar transforms. In Christian Boitet and Pete Whitelock, editors, *COLING-ACL*, pages 619–623. Morgan Kaufmann Publishers / ACL.
- Mark Johnson. 1998b. Pcfg models of linguistic tree representations. *Computational Linguistics*, 24(4):613–632.
- Mark Johnson. 2007. Transforming projective bilexical dependency grammars into efficiently-parsable cfgs with unfold-fold. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 168–175, Prague, Czech Republic, June. Association for Computational Linguistics.
- Hiroshi Kanayama, Yusuke Miyao, Takaaki Tanaka, Shinsuke Mori, Masayuki Asahara, and Sumire Uematsu. 2015. A draft proposal for universal dependencies japanese treebank (in japanese). In *In Proceedings of the 21st annual meeting for Gengo Shori Gakkai (The Association for Natural Language Processing)*.

- Daisuke Kawahara, Hongo Sadao, and Koiti Hasida. 2002. Construction of a japanese relevance-tagged corpus. In *In Proceedings of the 3rd International Conference on Language Resources and Evaluation*.
- Yasuhiro Kawata and Julia Bartels. 2000. Stylebook for the Japanese treebank in Verbmobil. In *Report 240*, Tübingen, Germany, September 29.
- Frank Keller. 2010. Cognitively plausible models of human language processing. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 60–67, Uppsala, Sweden, July. Association for Computational Linguistics.
- Kotaro Kitagawa and Kumiko Tanaka-Ishii. 2010. Tree-based deterministic dependency parsing — an application to nivre’s method —. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 189–193, Uppsala, Sweden, July. Association for Computational Linguistics.
- Dan Klein and Christopher D. Manning. 2002. A generative constituent-context model for improved grammar induction. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.
- Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan, July. Association for Computational Linguistics.
- Dan Klein and Christopher Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL’04), Main Volume*, pages 478–485, Barcelona, Spain, July.
- Arne Kohn and Wolfgang Menzel. 2014. Incremental predictive parsing with turboparser. In *Proceedings of the ACL 2014 Conference Short Papers*, Baltimore, USA, June. Association for Computational Linguistics.
- Lars Konieczny. 2000. Locality and parsing complexity. *Journal of Psycholinguistic Research*, 29(6):627–645.
- Terry Koo, Xavier Carreras, and Michael Collins. 2008. Simple semi-supervised dependency parsing. In *Proceedings of ACL-08: HLT*, pages 595–603, Columbus, Ohio, June. Association for Computational Linguistics.
- Matthias T. Kromann, Line Mikkelsen, and Stine Kern Lynge. 2004.
- Marco Kuhlmann, Carlos Gómez-Rodríguez, and Giorgio Satta. 2011. Dynamic programming algorithms for transition-based dependency parsers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 673–682, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Marco Kuhlmann. 2013. Mildly non-projective dependency grammar. *Computational Linguistics*, 39(2):355–387.

- Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 271–281, Baltimore, Maryland, June. Association for Computational Linguistics.
- Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. 2010. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1223–1233, Cambridge, MA, October. Association for Computational Linguistics.
- Tom Kwiatkowski, Sharon Goldwater, Luke Zettlemoyer, and Mark Steedman. 2012. A probabilistic model of syntactic and semantic acquisition from child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 234–244, Avignon, France, April. Association for Computational Linguistics.
- Tom Kwiatkowski, Eunsol Choi, Yoav Artzi, and Luke Zettlemoyer. 2013. Scaling semantic parsers with on-the-fly ontology matching. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1556, Seattle, Washington, USA, October. Association for Computational Linguistics.
- John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning, ICML '01*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.
- Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126 – 1177.
- Mike Lewis and Mark Steedman. 2014. A* ccg parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000, Doha, Qatar, October. Association for Computational Linguistics.
- Percy Liang, Michael Jordan, and Dan Klein. 2011. Learning dependency-based compositional semantics. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 590–599, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Dong C. Liu and Jorge Nocedal. 1989. On the limited memory bfgs method for large scale optimization. *MATHEMATICAL PROGRAMMING*, 45:503–528.
- Teresa Lynn, Jennifer Foster, Mark Dras, and Lamia Tounsi. 2014. Cross-lingual transfer parsing for low-resourced languages: An Irish case study. In *CLTW 2014*, Dublin, Ireland.
- Lluís Márquez M. Antónia Martí, Mariona Taulé and Manuel Bertran. 2007. CESS-ECE: A Multilingual and Multilevel Annotated Corpus. Available at <http://www.lsi.upc.edu/~mbertran/cess-ece/publications>.

- Christopher D. Manning and Bob Carpenter. 2000. Probabilistic parsing using left corner language models. In Harry Bunt and Anton Nijholt, editors, *Advances in Probabilistic and Other Parsing Technologies*, volume 16 of *Text, Speech and Language Technology*, pages 105–124. Springer Netherlands.
- Christopher D. Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts.
- Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *COMPUTATIONAL LINGUISTICS*, 19(2):313–330.
- David Mareček and Milan Straka. 2013. Stop-probability estimates computed on a large corpus improve unsupervised dependency parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 281–290, Sofia, Bulgaria, August. Association for Computational Linguistics.
- David Mareček and Zdeněk Žabokrtský. 2012. Exploiting reducibility in unsupervised dependency parsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 297–307, Jeju Island, Korea, July. Association for Computational Linguistics.
- Marie-Catherine De Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal stanford dependencies: a cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*, Reykjavik, Iceland, may.
- T. Matsuzaki, Y. Miyao, and J. Tsujii. 2007. Efficient hpsg parsing with supertagging and cfg-filtering. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI-07)*, pages 1671–1676, January.
- Ryan McDonald and Fernando Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, volume 6, pages 81–88.
- Ryan McDonald and Giorgio Satta. 2007. On the complexity of non-projective data-driven dependency parsing. In *Proceedings of the Tenth International Conference on Parsing Technologies*, pages 121–132, Prague, Czech Republic, June. Association for Computational Linguistics.
- Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT ’05, pages 523–530, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Ryan McDonald, Slav Petrov, and Keith Hall. 2011. Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the 2011 Conference on Empirical Methods in Natural*

- Language Processing*, pages 62–72, Edinburgh, Scotland, UK., July. Association for Computational Linguistics.
- Ryan McDonald, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Igor Mel'čuk. 1988. *Dependency Syntax: Theory and Practice*. State University of New York Press.
- George A. Miller and Noam Chomsky. 1963. Finitary models of language users. In D. Luce, editor, *Handbook of Mathematical Psychology*, pages 2–419. John Wiley & Sons.
- Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Nadia Mana, Fabio Pianesi, and Rodolfo Delmonte. 2003. Building the Italian syntactic-semantic treebank. In Anne Abeillé, editor, *Building and using Parsed Corpora*, Language and Speech series, pages 189–210, Dordrecht. Kluwer.
- Thomas Mueller, Richárd Farkas, Alex Judea, Helmut Schmid, and hinrich schuetze. 2014. Dependency parsing with latent refinements of part-of-speech tags. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 963–967, Doha, Qatar, October. Association for Computational Linguistics.
- K. Nakatani and E. Gibson. 2008. Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Linguistics*, 46(1):63–87.
- Kentaro Nakatani and Edward Gibson. 2010. An on-line study of japanese nesting complexity. *Cognitive Science*, 34(1):94–112.
- Tahira Naseem, Harr Chen, Regina Barzilay, and Mark Johnson. 2010. Using universal linguistic knowledge to guide grammar induction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1234–1244, Cambridge, MA, October. Association for Computational Linguistics.
- Tahira Naseem, Regina Barzilay, and Amir Globerson. 2012. Selective sharing for multilingual dependency parsing. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 629–637, Jeju Island, Korea, July. Association for Computational Linguistics.
- Mark-Jan Nederhof and Giorgio Satta. 2004a. Probabilistic parsing strategies. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume*, pages 542–549, Barcelona, Spain, July.

- Mark-Jan Nederhof and Giorgio Satta. 2004b. Tabular parsing. In Carlos Martín-Vide, Victor Mitran, and Gheorghe Păun, editors, *Formal Languages and Applications*, volume 148 of *Studies in Fuzziness and Soft Computing*, pages 529–549. Springer Berlin Heidelberg.
- Mark-Jan Nederhof. 1993. Generalized left-corner parsing. In *Proceedings of the Sixth Conference on European Chapter of the Association for Computational Linguistics*, EACL '93, pages 305–314, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Joakim Nivre and Daniel Fernández-González. 2014. Arc-eager parsing with the tree constraint. *Computational Linguistics*, 40(2):259–267.
- Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 99–106, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Joakim Nivre, Jens Nilsson, and Johan Hall. 2006. Talbanken05: A Swedish treebank with phrase structure and dependency annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC 2006)*, Genova, Italy. European Language Resources Association (ELRA).
- Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007a. The CoNLL 2007 shared task on dependency parsing. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932.
- Joakim Nivre, Johan Hall, Jens Nilsson, Atanas Chanev, Gülşen Eryiğit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007b. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95–135.
- Joakim Nivre. 2003. An efficient algorithm for projective dependency parsing. In *Proceedings of the 8th International Workshop on Parsing Technologies (IWPT)*, pages 149–160.
- Joakim Nivre. 2004. Incrementality in deterministic dependency parsing. In Frank Keller, Stephen Clark, Matthew Crocker, and Mark Steedman, editors, *Proceedings of the ACL Workshop Incremental Parsing: Bringing Engineering and Cognition Together*, pages 50–57, Barcelona, Spain, July. Association for Computational Linguistics.
- Joakim Nivre. 2008. Algorithms for deterministic incremental dependency parsing. *Computational Linguistics*, 34(4):513–553.
- Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 9041 of *Lecture Notes in Computer Science*, pages 3–16. Springer International Publishing.
- Hiroshi Noji and Yusuke Miyao. 2014. Left-corner transitions on dependency parsing. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2140–2150, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

- Hiroshi Noji and Yusuke Miyao. 2015. Left-corner parsing for dependency grammar. *Journal of Natural Language Processing*, 22(4), December.
- Hiroki Ouchi, Kevin Duh, and Yuji Matsumoto. 2014. Improving dependency parsers with supertags. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 154–158, Gothenburg, Sweden, April. Association for Computational Linguistics.
- John Pate and Sharon Goldwater. 2013. Unsupervised dependency parsing with acoustic cues. *Transactions of the Association for Computational Linguistics*, 1:63–74.
- Fernando Pereira and Yves Schabes. 1992. Inside-outside reestimation from partially bracketed corpora. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Newark, Delaware, USA, June. Association for Computational Linguistics.
- Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 433–440, Sydney, Australia, July. Association for Computational Linguistics.
- Emily Pitler, Sampath Kannan, and Mitchell Marcus. 2013. Finding optimal 1-endpoint-crossing trees. *Transactions of the Association for Computational Linguistics*, 1:13–24.
- Elias Ponvert, Jason Baldridge, and Katrin Erk. 2011. Simple unsupervised grammar induction from raw text with cascaded finite state models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Martin Popel, David Mareček, Jan Štěpánek, Daniel Zeman, and Zdeněk Žabokrtský. 2013. Coordination structures in dependency treebanks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 517–527, Sofia, Bulgaria, August. Association for Computational Linguistics.
- Prokopis Prokopidis, Elina Desipri, Maria Koutsombogera, Harris Papageorgiou, and Stelios Piperidis. 2005. Theoretical and practical issues in the construction of a Greek dependency treebank. In *In Proc. of the 4th Workshop on Treebanks and Linguistic Theories (TLT)*, pages 149–160.
- Philip Resnik. 1992. Left-corner parsing and psychological plausibility. In *COLING*, pages 191–197.
- Brian Edward Roark. 2001. *Robust Probabilistic Predictive Syntactic Processing: Motivations, Models, and Applications*. Ph.D. thesis, Providence, RI, USA. AAI3006783.
- D.J. Rosenkrantz and P.M. Lewis. 1970. Deterministic left corner parsing. In *Switching and Automata Theory, 1970., IEEE Conference Record of 11th Annual Symposium on*, pages 139–152, Oct.

- Francesco Sartorio, Giorgio Satta, and Joakim Nivre. 2013. A transition-based dependency parser using a dynamic parsing strategy. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 135–144, Sofia, Bulgaria, August. Association for Computational Linguistics.
- William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-coverage parsing using human-like memory constraints. *Computational Linguistics*, 36(1):1–30.
- Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 663–672, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Yoav Seginer. 2007. Fast unsupervised incremental parsing. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 384–391, Prague, Czech Republic, June. Association for Computational Linguistics.
- Mojgan Seraji. 2015. *Morphosyntactic Corpora and Tools for Persian*. PhD Thesis. *Studia Linguistica Upsaliensia* 16.
- Richard L. Lewis Shravan Vasishth. 2006. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794.
- Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.
- Kiril Simov and Petya Osenova. 2005. Extending the annotation of BulTreeBank: Phase 2. In *The Fourth Workshop on Treebanks and Linguistic Theories (TLT 2005)*, pages 173–184, Barcelona, December.
- Noah A. Smith and Jason Eisner. 2005. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL’05)*, pages 354–362, Ann Arbor, Michigan, June. Association for Computational Linguistics.
- Noah A. Smith and Jason Eisner. 2006. Annealing structural bias in multilingual weighted grammar induction. In *Proceedings of the International Conference on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL)*, pages 569–576, Sydney, July.
- Noah A. Smith. 2006. *Novel Estimation Methods for Unsupervised Discovery of Latent Structure in Natural Language Text*. Ph.D. thesis, Johns Hopkins University, Baltimore, MD, October.
- Otakar Smrž, Viktor Bieliký, Iveta Kouřilová, Jakub Kráčmar, Jan Hajič, and Petr Zemánek. 2008. Prague Arabic dependency treebank: A word on the million words. In *Proceedings of the*

- Workshop on Arabic and Local Languages (LREC 2008)*, pages 16–23, Marrakech, Morocco. European Language Resources Association.
- Anders Søgaard. 2012. Two baselines for unsupervised dependency parsing. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, pages 81–83, Montréal, Canada, June. Association for Computational Linguistics.
- Valentin I. Spitzkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2010a. From baby steps to leapfrog: How “less is more” in unsupervised dependency parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 751–759, Los Angeles, California, June. Association for Computational Linguistics.
- Valentin I. Spitzkovsky, Hiyan Alshawi, Daniel Jurafsky, and Christopher D. Manning. 2010b. Viterbi training improves unsupervised dependency parsing. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, pages 9–17, Uppsala, Sweden, July. Association for Computational Linguistics.
- Valentin I. Spitzkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2012. Three dependency-and-boundary models for grammar induction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 688–698, Jeju Island, Korea, July. Association for Computational Linguistics.
- Valentin I. Spitzkovsky, Hiyan Alshawi, and Daniel Jurafsky. 2013. Breaking out of local optima with count transforms and model recombination: A study in grammar induction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1983–1995, Seattle, Washington, USA, October. Association for Computational Linguistics.
- Mark Steedman. 2000. *The Syntactic process*. Language, speech, and communication. MIT Press, Cambridge (Mass.), London. A Bradford book.
- Andreas Stolcke. 1995. An efficient probabilistic context-free parsing algorithm that computes prefix probabilities. *Comput. Linguist.*, 21(2):165–201, June.
- Oscar Täckström, Ryan McDonald, and Joakim Nivre. 2013. Target language adaptation of discriminative transfer parsers. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1061–1071, Atlanta, Georgia, June. Association for Computational Linguistics.
- L Tesnière. 1959. *Elements de syntaxe structurale*. Editions Klincksieck.
- Leonoor van der Beek, Gosse Bouma, Jan Daciuk, Tanja Gaustad, Robert Malouf, Gertjan van Noord, Robbert Prins, and Begoña Villada. 2002. Chapter 5. the Alpino dependency treebank. In *Algorithms for Linguistic Processing NWO PIONIER Progress Report*, Groningen, The Netherlands.

- Marten van Schijndel and William Schuler. 2013. An analysis of frequency- and memory-based processing costs. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 95–105, Atlanta, Georgia, June. Association for Computational Linguistics.
- Marten van Schijndel, Andy Exley, and William Schuler. 2013. A model of language processing as hierarchic sequential prediction. *Topics in Cognitive Science*, 5(3):522–540.
- Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. 2005. Maximum margin clustering. In *Advances in Neural Information Processing Systems 17*, pages 1537–1544. MIT Press.
- H. Yamada and Y. Matsumoto. 2003. Statistical Dependency Analysis with Support Vector machines. In *The 8th International Workshop of Parsing Technologies (IWPT2003)*.
- Daniel Zeman, Ondřej Dušek, David Mareček, Martin Popel, Loganathan Ramasamy, Jan Štěpánek, Zdeněk Žabokrtský, and Jan Hajič. 2014. Hamletd: Harmonized multi-language dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.
- Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the 21st Conference on Uncertainty in AI*, pages 658–666.
- Yue Zhang and Stephen Clark. 2008. A tale of two parsers: Investigating and combining graph-based and transition-based dependency parsing. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 562–571, Honolulu, Hawaii, October. Association for Computational Linguistics.
- Yue Zhang and Joakim Nivre. 2011. Transition-based dependency parsing with rich non-local features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 188–193, Portland, Oregon, USA, June. Association for Computational Linguistics.
- Arnold M Zwicky. 1993. Heads, bases and functors. *Heads in Grammatical Theory*, page 292.